Foundations for Quality Management of Scientific Data Products

NICOLE M. RADZIWILL, NATIONAL RADIO ASTRONOMY OBSERVATORY

The costs of making incorrect scientific inferences based on faulty data can be substantial and far-reaching: errors can be subtle, inappropriate conclusions can go unchallenged for years in the literature, and follow-on research may be critically jeopardized. Because most scientific research in the United States is federally funded, the propagation of errors through research studies imbues high costs to taxpayers over time; errors in scientific conclusions, and any technologies based on them, will require rework at some point in the future. Better scientific data quality means more accurate conclusions can be made more quickly, and benefits can be realized by society more readily. To improve scientific data quality, and provide continuous quality assessment and management, the nature of scientific data and the processes that produce it must be articulated.

The purpose of this research is to provide a conceptual foundation for the management of data quality as it applies to scientific data products, specifically those generated by the large-scale instrumentation and facilities that will populate the data centers of the future. Definitions for data product and data quality tailored to the context of scientific decision making are proposed, given two typical scenarios: 1) collecting observational data, and 2) performing archive-based research. Two relevant extensions to the total quality management (TQM) philosophy, total information quality management (TIQM), and total data quality management (TDQM) are then examined to determine if the management of scientific data quality differs from the management of other data or information. Recommendations for planning, assessment/assurance, control, and continuous improvement are proposed, focusing on designing quality into the production process rather than relying on mass inspection.

Key words: data center, data management, data quality, scientific data, semiotics, TDQM, TIQM, TQM

INTRODUCTION

When producing scientific data products (SDPs), managing for quality is particularly important because use of the data may pose broad impacts to the efficiency of an organization, to future scientific investigation, or to society. Previous research has outlined or attempted to quantify some of these impacts:

- The Institute for the Study of Society and Environment, as part of its "Societal Impacts of Weather" project, estimates that $300 million in economic losses per week can be attributed to extreme weather (ISSE 2005). Better weather forecasts can save up to $250,000 a day per oil drilling rig that is affected. Every avoided flight cancellation saves approximately $40,000, and each avoided diverted flight saves an estimated $150,000. Bad forecasts also cost money; a grounded airplane (or one that is inappropriately grounded) will result in a loss of revenue for the carrier. Each of these scenarios requires high-quality raw data, combined in accurate and precise ways to yield intelligible model forecasts of weather conditions.
- Human safety can depend on high-quality science data, because with better detection of environmental hazards, future hazardous situations and catastrophic events can be predicted more accurately. Consider the December 26, 2004, tsunami in the Indian Ocean, which may have been predicted if high-quality raw and derived geophysical data were available from that region at the time.
- Inefficiencies in operations often result from poor quality data, or even a lack of data quality information. Access to major scientific facilities (for

example, telescopes) is often either over-allocated or under-allocated when a research proposal does not accurately address the time requirements to meet a data quality goal (O'Neil 2005).

- One of the strategic goals of the National Science Foundation (NSF) is to enable cross-fertilization of scientific ideas, allowing nonspecialists in one field to explore datasets in another field that may be outside their realm of expertise. These crossover researchers will not be as sensitive to subtle variations in the quality of data products, and so to reach this state of operations, the quality of datasets must be managed more stringently in the future.

The defining characteristic underlying solutions to all of the above is the fact that modern scientific techniques are capable of producing higher volumes of dynamic, complex information at faster rates than ever before. To provide the solutions for these new challenges, effective scientific data management is critical, and as a result has been strongly funded by NSF for nearly a decade. Although there are several definitions of scientific data management, software and hardware aspects of all of the following are typically considered: archival, presentation, preservation, retrieval, and investigation of data; mass data storage, mass manipulation, and acquisition of data from distributed and possibly heterogeneous sources; movement and transport of large quantities of data; high-resolution visualization capabilities; and data assimilation (which includes collecting, combining, repairing, assessing/verifying, normalizing, and adjusting to different reference frames if required). Projects such as TeraGrid (http://www.teragrid.org), a "multiyear effort to build and deploy the world's largest, most comprehensive, distributed infrastructure for open scientific research," aim to provide the basic and applied research necessary to enable the aforementioned capabilities.

Complex datasets that require effective scientific data management will largely be those produced by next-generation science instrumentation. Domestic initiatives include the National Ecological Observatory Network (NEON) to improve environmental forecasting, the Rare Symmetry Violating Processes (RSVP) installation for explorations of particle physics and dark matter, the EarthScope network of geophysical instrumentation, and the IceCube Neutrino Observatory. International collaborations include the Atacama Large Millimeter Array (ALMA) telescope, which will provide "astronomy's most versatile imaging instrument" (NSF 2006). The United States will also be contributing to a 10-year effort to build the Global Earth Observing System of Systems (GEOSS), which includes a worldwide tsunami detection facility, and has funded an exploratory research program for the U.S. National Virtual Observatory to promote data-intensive astronomical research. Most, if not all, of these facilities will be affiliated with a science data center.

The science data center is a powerful emerging organizational trend that can be expected to dominate the landscape of science management within the next two decades. Pioneering research at Microsoft's Advanced Technology Division on the topic of scientific data management explains that "science centers that curate and serve science data are emerging around next-generation science instruments" (Gray et al. 2005). These same individuals, however, report that because "these new instruments have extraordinary precision, the data quality is rapidly improving." Because data quality is not limited to the dimension of precision, improved quality cannot be obtained solely by increasing the sensitivity of the instrumentation. This simplistic viewpoint fails to acknowledge that quality products are the result of a process that engenders quality in all its stages.

In fact, there are many similarities between the nature and competitive environment of process industries (for example, chemicals, pharmaceuticals) and scientific data production from large facilities. Like the process industries, scientific data collection and generation is often impacted by environmental conditions, some of which can be controlled. Competitive advantage at all these facilities is gained by custom engineering of advanced components, which often impedes collaboration as well as the standardization of processes, leading to a natural resistance to implement quality management practices. The outputs in both cases are raw materials that are intended for other assembly lines or information processing programs (Spivak and Brenner 1998).

Notably absent from the concept of scientific data management as it stands today is the pursuit of product and process quality applied to the SDP. Because the emergence of science data centers will underscore the need for high-quality SDPs, quality management in this field will become a necessity instead of an afterthought. Ten years from now opportunities will abound for quality professionals to apply their skills to the scientific domain. To do so, however, the theoretical foundations for providing practical solutions to data quality issues in this realm must be established. The current study aims to begin the process of defining these foundations.

DATA QUALITY AT THE NATIONAL RADIO ASTRONOMY OBSERVATORY

To explore the concept of scientific data quality it is critical to review representative examples and examine the drivers for quality in these cases. Over the course of a year and a half, the nature of scientific data and data quality at the National Radio Astronomy Observatory's Robert C. Byrd Green Bank Telescope (GBT) in Green Bank, W.Va., was examined as part of a small-scale software quality diagnostics project. For comparisons, the author drew on previous work with meteorological numerical weather prediction (NWP) models at the Institute for Atmospheric Sciences in Rapid City, S.D., and experience developing a new meteorological observing system for water vapor remote sensing using Global Positioning System (GPS) signals at the National Oceanic and Atmospheric Administration (NOAA) in Boulder, Colo. The analysis focused on GBT data production, which is the most complex of the cases.

The GBT is a 100-meter single-dish radio telescope with a unique unblocked aperture design and fully adjustable surface panels. These two features contribute to making the GBT one of the most sensitive and precise single-dish telescopes in the world: reflection and diffraction that typically compromise the beam pattern of a radio telescope are reduced, and the dish can be adjusted to maintain a near-perfect parabolic shape regardless of the orientation of the 17 million pound structure. Radio waves are focused by the dish onto one of the telescope's receivers, which operate at frequencies from approximately 200 MHz to 50 GHz. Development is in progress to increase the upper range through 115 GHz. After the signal is received, it is amplified and mixed several times before arriving at a special-purpose signal-processing device called a backend. The astronomer observing with the GBT chooses one or more of these backend devices to match the scientific intent of his or her research program. Connecting these devices to achieve the desired intent requires setting hundreds of control system parameters for each observation; errors in any of these steps reduce the quality of resultant data products.

The quality of the final data product depends upon the health of the hardware components, as well as the computing systems, computing infrastructure, and software modules at the time the observation is collected. Because the GBT is a pseudo real-time system, even millisecond-scale delays in communication between devices or software modules can negatively impact raw data quality. The quality of any data derived from the raw data is, understandably, dependent upon the quality of the raw data that are input. A similar philosophy is held by the European Southern Observatory (ESO) in Germany, where the Quality Control and Data Flow Operations Division monitors "the quality of the raw data, the quality of the products and of the product creation process, and the performance of the instrument components involved" (Hanuschik et al. 2002).

NRAO Green Bank (GB) competes for funding with other research facilities. To enhance competitiveness, the operational strategy must continually minimize the time to market for delivering new scientific capabilities from the instrument, while shortening the cycle time to identify and implement changes to the instrument's hardware and software to improve data quality. These quality goals are just as significant as the primary driver for quality improvement efforts, which is to produce datasets capable of satisfying the scientific intent of the astronomer who observed the data.

The data production process for the GBT is similar to build-to-order or mass customization approaches, where the end product is highly dependent upon the consumer's input parameters. There are three ways that traditional manufacturers can customize their products: 1) modular customization, where building blocks are joined together to form complete products; 2) adjustable customization, where different configurations or rules applied to a product yield essentially different products; and 3) dimensional customization, which produces essentially unique products. Examples of the latter include tailoring of clothing and custom machining parts used in subassemblies (Anderson 2002). The scientific data produced by the GBT is dimensionally customized, a characteristic shared with the data products from other telescopes and many other scientific data production processes as well. The key difference between the GBT and other telescopes (such as interferometers, or the single-dish telescope at the Arecibo Observatory in Puerto Rico) is that the realm of possible customizations is much larger because of the extensive flexibility designed into the GBT.

Because of this dimensional customization, it is not possible to select representative data samples from the aggregated output of the GBT over time and compare them "apples to apples," or through the statistical analysis of batches selected from the solution space of the data products that have been produced, for use in quality-related improvement activities. As a result, it is difficult for many to identify how quality control can be achieved without mass inspection. Mass inspection attempts to control the product instead of the process; inspection won't find all defects and quality problems determined by inspection require rework. Prevention of quality problems will yield better results, and so a key requirement for a data quality management framework is to be able to prevent quality issues from becoming evident in the data products to be consumed by the researchers who are collecting the data.

To manage data quality at NRAO, a planning methodology must be available that takes into consideration quality characterization and assurance, quality control, and continuous improvement. There must be methods for assessing the quality of an observational dataset as it is being produced with respect to its intended use. This is important because the telescope itself represents a dynamic decision environment; the use of the instrument can be adjusted in real-time to follow changes in the observer's program, or to capture new scientific opportunities as they emerge during an observation. By doing so, this also helps to optimize the value delivered by the instrument over its lifetime. Metadata describing the intermediary data products and the collection process should also be archived and indexed along with the data products for future inclusion in a science data center. This is a consideration because future researchers will need to assess the comparative quality of archive datasets, so that they may choose between otherwise similar historical datasets to support their science. Continuous quality improvement is significant because given a method to manage and continuously improve the quality of NRAO data products, the productivity of the instrument will be enhanced. Furthermore, this is a critical step toward the goal of intelligent processing of raw and derived data products. That is, observational and control system software raises flags if output anywhere in the production process does not conform to data quality objectives, and the process is autonomous or self-repairing wherever possible.

Although this research has focused on providing quality astronomical raw data and derived data products (including images, spectra, and n-dimensional aggregations of data called data cubes), other data products conform to the structures and quality management models investigated. These include data collected from meteorological observation systems, assimilated data used as input to weather prediction models, and NWP model output. Other datasets with similar characteristics include medical spectroscopy, imaging results, and input/output from engineering simulations (such as finite element models), although these were not specifically addressed in this study.

QUALITY MANAGEMENT OF DATA AND DATA PRODUCTS

Poor data quality limits its usefulness. In scientific research, data products are a means to an end; namely, the identification, publication, and dissemination of scientific results. However, for suppliers who provide data to the researcher-consumer, the data product is the primary reason for the existence of the production facility. That the data production facility might also design, construct, and manage sophisticated instrumentation is secondary to the researcher, who is interested primarily in the data products to be analyzed. Since the early 1990s, there have been several studies of data quality, information quality, and generating information as a product that results from a manufacturing process. Several studies have examined the questions associated with data quality diagnosis and management, and these are summarized later. To date, however, no research has been specifically geared to the issues that arise in the production and quality management of SDPs.

Wang and Strong (1996) classified 15 attributes in four categories: intrinsic data quality, contextual data quality, representational data quality, and quality of accessibility. These attributes were further explored in *Data Quality for the Information Age* (Redman 1996) and refined once again in *Enterprise Knowledge Management: The Data Quality Approach* (Loshin 2001). Loshin's synthesis of the quality attributes as they apply to data are summarized in Figure 1.

Researchers have attempted to characterize data quality in terms of quality problems perceived by users. For example, Becker (1998) identified seven categories of data quality problems: data corruption, different meanings in historical and current data, multiple data definitions, missing data, hidden data, inappropriate granularity of data, and violations of data integrity rules. Other investigators have examined the root causes for poor data quality. In one study, these were determined to be process problems, system problems, policy or procedure problems, and data design issues (Cykana and Stern 1996). Another study revealed that the key contributor to poor data quality is process issues, suggesting that an understanding of the process by which data are generated is essential to understanding and improving data quality (Dvir and Evans 1996).

Most applied data quality research to date has focused on large corporate databases or enterprise information stores, where the goal of data mining is to yield high-quality decision-making information for business purposes. In order for the business decisions to be sound, the data must be of a certain quality level. In these information stores, defective data can be corrected, increasing the utility of future decision making. Typical problems that impact these data stores include inappropriate duplication, the presence of inaccurate information, and variations on the same referenced entity (for example, Street, St., and Str.). The main purpose of ensuring data quality in these circumstances is to enable effective decision support from the data warehouse. Many of these studies have been done by premier consulting firms. For an example, refer to Tsien (2004). These studies do not consider data quality issues as they apply to an inventory of information products, as in the case of the scientific data archive.

Data quality management has also been an active topic of research, which to date has been based upon the total quality management (TQM) philosophy. Kujala and Lillrank (2004) noted that TQM has developed as a practice-oriented discipline, and was not founded on a "theoretically solid or empirically validated framework," suggesting that data quality research also requires a more substantive basis. TQM-based quality management frameworks developed for data quality include total information quality management (TIQM), (English 1999) and the "guidelines" approach to data quality management coordination (Loshin 2004). Data quality costs have also been explored, ranging from a simplistic assessment of costs related to prevention, detection, and repair of quality issues (Kim and Choi 2003), to the development of the cost of poor data quality (COPDQ) metric (English 2004). The total data quality management (TDQM) research effort headed by Richard Wang at the Massachusetts Institute of Technology (MIT) seeks to "create a theory of data quality based on reference disciplines such as computer science, the study of organizational behavior, statistics, accounting, and the total quality management field… which may serve as a foundation for other research contexts where the quality of information is an important component" (MIT 2005). This program's research results include a framework for process mapping called IPMAP that treats data production as a manufacturing process, and posits the notion of a virtual business environment that is characterized by dynamic decision making and thus requires efficient

Figure 1 Loshin's dimensions of data quality described.

Data Models

- **Clarity of definition**—unambiguous labels, distinctive and representative naming of objects, attributes, and relations.
- **Comprehensiveness**—suitability of model for use by current applications, and adaptability of model to future uses.
- **Flexibility**—ability of data model to be adapted to future usage scenarios.
- **Robustness**—ease of adaptation of the model to future usage scenarios.
- **Essentialness**—not including extraneous information; including fundamental information from which other quantities can be easily derived.
- **Attribute granularity**—selecting the appropriate number of objects to represent a concept.
- **Precision of Domains**—allocating appropriate precision (ex. data types, sizes) to a data attribute.
- **Homogeneity**—not hiding business rules or the simplifying assumptions used by applications in the data.
- **Naturalness**—not "overloading" data values with metadata or supplemental descriptive characteristics.
- **Identifiability**—ensuring that there is a way to distinguish between similar entities (ex. using primary keys).
- **Obtainability**—is it legal, feasible, and/or appropriate to collect, persist, and use a particular data value?
- **Relevance**—are all the stored attributes useful by applications, or are there plans to use them?
- **Simplicity**—the model does not encourage complications in the applications that will use the data.
- **Semantic consistency**—ensure that the meanings and names of objects within a dataset are consistent, that one term, not several, are used to refer to the same physical concept.
- **Structural consistency**—ensure that ideas are uniformly referenced, including use of standard units.

Data Values

- **Accuracy**—the level to which stored data agree with accepted sources of "correct" information.
- **Null values**—the absence of information can provide valuable insight into obtainability/relevance problems.
- **Completeness**—how well the expectation that certain fields will be appropriately populated is met.
- **Consistency**—data values in one set being logically consistent with those in another related set.
- **Currency/Timeliness**—degree to which information represents the up-to-date state of what is being modeled.

Information Domains

- **Enterprise agreement of usage**—communicating in terms of nomenclature to which all have conformed.
- **Stewardship**—ensuring that responsibility for maintaining integrity and currency of attributes is assigned.
- **Ubiquity**—encouraging sharing of data resources and standardization of data use across applications.

Data Presentation

- **Appropriateness**—how well the format and content of data satisfies the users' needs.
- **Correct interpretation**—how comprehensive the provided information is, so that users can make accurate inferences.
- **Flexibility**—adaptability of system to changes in represented information or requirements for the use of that information.
- **Format precision**—ensuring that the stored, displayed, and leveraged data instances contain the required granularity of information for the application for which they are used.
- **Portability**—how well the capability to move applications from one platform to another has been preserved.
- **Representation consistency**—whether instances of data effectively implement the consistency demanded by the model.
- **Representation of null values**—representing all null or missing values equivalently.
- **Use of storage**—ensuring that the storage of the data effectively uses the media upon which it resides.

Information Policy

- **Accessibility**—degree of ease of access, and breadth of access, to information.
- **Metadata**—ensuring that metadata are not only defined but also meet required dimensions of data quality.
- **Privacy and security**—ensuring that an approach is in place to display information selectively, and also to protect the integrity of the data and metadata themselves.
- **Redundancy**—being cognizant of where and why repetitive data are useful, and managing the inflow of data accordingly.
- **Unit cost**—understanding the costs associated with persisting and maintaining information, and being parsimonious where possible, so that the total cost of ownership of data is reduced or minimized.

Figure 2 Summary and description of scientific data types. Levels 0 through 4 can be treated as SDPs if they are offered to the customer for review or for scientific analysis, and thus require a subjective determination of fitness for use.

Domain	Data Type	NASA/EOS Terminology	Description
Instrument	Device monitoring data	N/A (Level-1)	Produced by instrumentation, typically not preprocessed, and typically not stored as part of raw data products. Useful for preparing trending data to detect emerging instrument failures, and providing operational responses to other failures to dynamically improve the potential for producing quality data products.
Instrument	Raw data	Level 0	Produced by instrumentation. May be subject to limited preprocessing in firmware (for example, autocorrelation spectrometers).
Instrument	Calibrated data	Level 1	Produced by removing instrumental and environmental effects. EOS breaks this down into Levels 1A (raw data appended with annotations and calibration information) and Level 1B (raw data processed to calibrated data).
Science	Derived data	Level 2	Produced by combining calibrated data with other calibrated data, or with other derived data, according to processes, techniques, or algorithms. Scientific analysis can take place at this level or any higher level.
Science	Assimilated data	Level 3	Produced by gridding, resampling, and/or changing the frame of reference for derived data.
Science	Model data	Level 4	Produced by applying one or more mathematical, physical, or stochastic models to collections of assimilated and derived data products.

data quality management (Shankaranarayan, Ziad, and Wang 2003).

Though TDQM treats data as an information product, its comparison of product manufacturing and information manufacturing assumes that raw data are the input to an information system (the process) to yield information products (Wang 1998). The mission of the scientific data production facility, however, is to produce the raw data that are then processed to form other scientific data products. Examining the issues associated with the production of raw data has not been an aim of the TDQM research program, and neither TIQM nor TDQM consider the role of data quality in an archive, or the process of continually improving the quality of archived products based on new knowledge.

Although significant work has been pursued regarding data quality and its management, no research has evaluated the nature of the SDP, or data quality or management as it applies to this domain.

WHAT IS A SCIENTIFIC DATA PRODUCT?

To understand what defines an SDP, one must first examine the types of scientific data that are produced, quality objectives for each of these types, and how the types combine to form products. In a complex instrumentation system, such as any next-generation research equipment facility, component instruments collect data. Once combined, preprocessed or aggregated, these datasets become "raw data." When instrumental and environmental effects are removed, this becomes "calibrated data." When calibrated data are transformed or combined by the application of algorithms, products may result that can be analyzed in their own right to yield scientific results. The term for this is "derived data." Beyond derived data, one may automatically or manually remove what is subjectively

Figure 3 Characteristics of science data products (SDPs).

Characteristics of Scientific Data Products (SDPs)

- Credibility of the facility producing the data is established by the quality of the data themselves (or lack thereof).
- Multiple, nested, interdependent combinations of raw, calibrated, derived, and assimilated data types.
- Product requires application of knowledge and insight to generate scientific conclusions.
- Usually depend on the performance and complex interplay of a system of hardware, software, and algorithms executed at a particular time.
- Quality is strongly dependent upon integrity of algorithms and production process, which are usually implemented in software.
- Quality of the product depends not only on the state of the system performing the observation, but also on the state of the object(s) or system(s) being observed.
- Continual refinement (for example, data cleansing) is often not performed, because the data product typically represents an observation at a given point in time.
- Data cleaning has limited applicability because there is often not a way to know what the "correct" values are.
- Repairing databases for duplicates or multiple data representations is typically not a concern.
- Data entropy is not a concern; unlike addresses, phone numbers, or banking information, data do not tend to become incorrect over time—the purpose of collecting the science data is to capture the state of the observed system at that time.
- Because continuous learning is the goal of science, and this tends to reveal new insights about how to improve data quality, incremental improvements must be made to produce better products; continuous improvement is thus a required aspect of the production process.
- When new knowledge is uncovered, old datasets may or may not be reprocessed (unless required by a particular research project) whereas in a data warehouse, reprocessing would be critical for data integrity.
- Spans multiple data representations (arrays, images, events, data cubes, or objects such as proteins or models).
- Quality depends upon suitability of algorithms chosen to satisfy scientific intent.
- Feature extraction can be expected to take place to produce new objects that support novel types of analysis.
- May be available for many scales of physical representation, and there may be interactions between the scales (for example, derived GPS data products as input to NWP; single-dish data products as input to astronomical interferometry experiments).
- Requires metadata so consumers can make judgments about validity and applicability of results.
- Data quality improves with time as new insights are gained regarding what constitutes quality in the product itself or within a stage of the production process.

©2006, ASQ

determined to be bad data, and may choose to regrid or otherwise normalize the data; the result is "assimilated data." In some disciplines, these assimilated data can be used as input to complex models, yielding model output data.

A summary of these data classifications, which draws from NASA's level 0 to level 2 data product model (described at http://observer.gsfc.nasa.gov/sec3/ProductLevels.html), and the NASA Earth Observing System (EOS), which augments the NASA model with levels 3 and 4 (described at http://www.srl.caltech.edu/ACE/ASC/level1/dpl_def.htm), is shown in Figure 2. Considered jointly, this model accurately characterizes the spectrum of data classifications that may be encountered at NRAO Green Bank. A unidirectional dataflow is not implied by this categorization. Derived and assimilated SDPs, for example, can be produced by multiple, nested combinations of both raw and derived data. Data type and NASA/EOS terminology are used interchangeably throughout the remainder of this study.

Two additions have been made to increase the utility of this categorization in a quality management model. The device monitoring data type was added because the health of devices producing raw data is critical to producing raw data. Also, the categories were associated with the "instrument domain" or the "science domain." At NRAO, the execution of an observation and the production of data both fall in the instrument domain. The reduction and analysis of data are both in the science domain, since the approaches and algorithms used can be the same from instrument to instrument, and are not necessarily dependent upon the characteristics of the instrument itself (Tody 2004). The implication of this addition is that the data models between the two domains must be separate to ensure data quality across the data categories; processes in the science domain do not require knowledge of the instrument domain, and processes in the instrument domain should not attempt to apply processing that requires knowledge from the science domain.

What, then, constitutes an SDP? Because the customer provides the ultimate assessment of quality, whether or not the dataset is fit for his or her intended use, any dataset that can be offered to the customer as an intermediary data product or for scientific analysis should be considered an SDP. SDPs can fall in any of the categories from level 0 through level 4. The resultant quality of the SDP will depend upon the quality of all the ingredients in its production process, which can come from its own level or any level above. The characteristics of an SDP that distinguish it from other types of data or information products are summarized in Figure 3.

THE NATURE OF SCIENTIFIC DATA

The quality of an SDP can be summarized as its suitability to satisfy a desired scientific intent, which is a variation on J. M. Juran's "fitness for use" qualification (Juran and Gryna 1988). Fitness for use is also the primary criterion for judging data quality according to TDQM (Wang 1998). Note that in order for an SDP to be fit for use, it must meet several of the data quality attributes described in Figure 1, such as accessibility, appropriateness, currency, and timeliness. Since the SDP is only a means to an end, namely the generation and dissemination of sound scientific conclusions, its quality must be assessed in the context of a decision process.

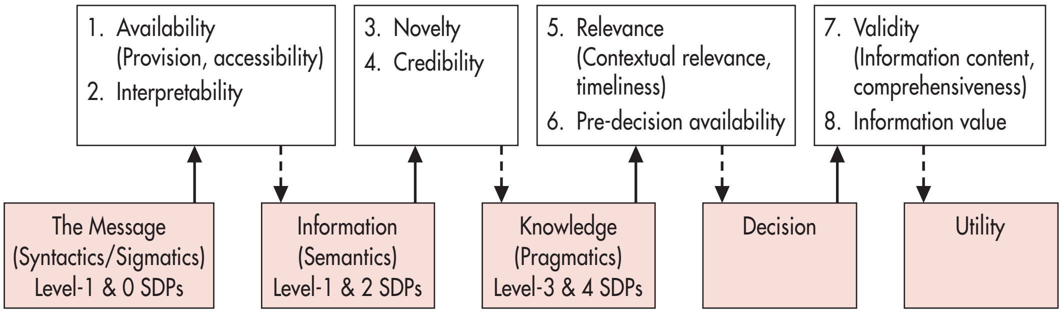
Expanding on the concept of data quality dimensions, some investigators (Graefe 2003; Price and Shanks 2005) have turned to the fundamentals of semiotics to explore information quality as it relates to the decision process. Semiotics is the theory and study of signs and symbols, integrating syntactics (basic relationships between signs and symbols), sigmatics (the relationship of data to the signs and symbols it is represented by), semantics (deriving meaning from data), and pragmatics (generating knowledge from meaningful interpretations of data).

Because SDP quality is so strongly dependent upon the integrity of the production process, and the production process follows the progression of the semiotic branches, Graefe's criteria of information quality in the decision process can be associated with the various levels of SDP, as illustrated in Figure 4.

This representation is particularly appropriate because it:

- Integrates established data quality attributes with a model of the decision process and the description of scientific data types outlined earlier.
- Appropriately identifies the need for novelty (new messages that are capable of influencing existing knowledge and impending decisions) in both level 1 and level 2 SDPs, which span both the instrument and the science domains.
- Accurately captures the goal of quality management for SDPs, which is to increase the utility of the product within the context of the scientific decision-making process.

Figure 4 Data quality criteria for SDPs in the context of a decision process (adapted from Graefe 2003).



©2006, ASQ

With respect to Loshin's classification of data quality dimensions outlined in Figure 1, the data quality criteria for SDPs assume that data quality of data models has been achieved, encompasses data quality of data values within the "interpretability" dimension, and addresses data quality of data presentation within the "credibility" dimension, since these elements are required to produce credibility. To generate utility from the decision process, quality of information domains and quality of information policy must be integrated into the quality management approach.

CONSIDERATIONS FOR A QUALITY MANAGEMENT MODEL

Poor data come from a combination of factors, including human error in the design of the experiment, inconsistent or inaccurate data generation processes, or faulty instrumentation. However, in the case of SDPs produced by a facility dedicated to the generation and processing of scientific information products, even good quality data from the instrument may be unusable (and thus of poor quality) for a different, future experiment. Thus, even if processes are managed well, poor data quality will still exist. Methods to assess data quality with respect to an experiment's unique objectives are thus a critical component of a quality management system.

Critical examination of SDPs and what constitutes SDP quality has yielded a multilayered definition for various classifications of SDP, and models for quality diagnostics and SDP production rooted in semiotics. The goals the author aims to accomplish by managing scientific data quality, however, are: 1) to detect failures in the production process and fix them before they impact the raw data; 2) to be able to identify issues in raw data prior to producing and analyzing it to produce higher-level SDPs; and 3) to address root causes of problems rather than symptoms that exhibit themselves in data quality at the various levels of SDP. The intent underlying these goals is to derive optimal utility from the decisions that can be made using the knowledge.

It is this void that a quality management model can fulfill. There are many well-established management methods for achieving total quality in an organization, including TQM, Six Sigma, TQC, Kaizen, and techniques such as quality function deployment (QFD). Three common characteristics unify these methods (Evans and Lindsay 2005):

- A focus on customers and stakeholders
- Participation and teamwork by everyone in the organization
- A process focus supported by continuous improvement and learning

All methods provide strategies for quality planning, assurance, control, and improvement. For production of high-quality SDPs, the process must include feedback into the data production process, and a commitment to identifying and addressing data quality issues for all new capabilities to be added (which includes generating requirements for production and analysis software).

SDPs are, primarily, information products. As a result, principles of information quality can be applied to yield product quality. TIQM integrates W. Edwards Deming's 14 Points of Quality with elements from Kaizen and principles conceived by Philip Crosby and J. M. Juran. The methodology relates these points and principles to the special concerns of managing data and information quality, and specifies six processes, which are outlined in Figure 5.

P1 and P3 are applicable to the planning stages of quality management, whereas P2 applies to quality assessment and assurance. P4 relates to control, and P5 and P6 map to continuous improvement. Each of the four high-level elements required for scientific data quality management are present, but reengineering and correcting data (process P4) requires mass inspection in an environment where the products are produced by dimensional customization.

There are several distinctions that are not specifically addressed within TIQM, yet are required to effectively manage the quality of SDPs at the various levels. First, TIQM does not address the importance of normal versus assignable cause variation in the

Figure 5 Description of the six processes of total information quality management (TIQM).

P1: Assess data definition & information architecture quality. Assess the quality of data models, database and data warehouse designs, and stability and consistency of systems architectures supporting the quality of data in the enterprise.

P2: Assess information quality. Identify how well-generated data conform to the dimensions of data quality that are deemed important within the organization.

P3: Measure nonquality information and risks. Establish a quality costs program, where costs associated with prevention, appraisal, internal failure, and external failure are measured and tracked.

P4: Reengineer and correct data. When faulty data are produced, repair those data. Conduct joint P5 initiatives to ensure that the cause of failure is determined, and the process improved to eliminate the root cause of the fault.

P5: Improve information process quality. Deliver value to the organization by repairing processes that are error prone, solving upstream issues before they propagate downstream and lead to rework.

P6: Establish the information quality environment. Apply TIQM's 14 points of information quality to cultivate a quality culture.

©2006, ASQ

determination of data quality, which is critical at all stages of the SDP production process. Because data products are ultimately intended for an archive, process quality problems must be solved readily, otherwise a larger volume of lower quality data will be produced and archived in the meantime. Also, the reengineering and correction of faulty data required in P4 is often not practical, or possible, in science—particularly when observational systems are generating the SDPs, and the new algorithms that would enable the data cleansing take a long time to emerge. Repairing the data should be considered a much lower priority than finding ways to automate the repair of the process, thus yielding a self-repairing system. Additionally, a quality management model must provide a subjective means of assessing and controlling data quality in addition to an objective means. These considerations, when integrated into TIQM, formed a customized, more detailed operationalization of TIQM for the particular case of data from the GBT (see Appendix 1). While this describes one example of how the TIQM processes can be interpreted for application to scientific data management, Appendix 2 outlines the 14 points of Information Quality that can be applied as is to develop an information quality culture as part of an organization's continuous improvement efforts. Although the content of Appendix 1 represents preliminary research that has not yet been empirically validated to form a general quality management model for scientific data production, its development was based on the definitions and structures described in previous sections.

CONCLUSIONS AND FUTURE WORK

A conceptualization or taxonomy is not intended to provide prescriptive guidelines and cannot be empirically validated; its value comes from its utility and effectiveness as a tool for gaining greater insights. Using the classifications developed herein, it should be possible to establish the theoretical foundations for quality management in the sciences, as well as to identify and develop techniques for quality improvement based on the extensive body of knowledge that has already been applied to other disciplines. Understanding the nature of quality as it applies to scientific data enables one to focus on the most applicable issues for future research and development. As a result, it may be impractical for the quality manager to generalize these findings for application in other environments. However, the similarity between process industries and the process of generating scientific data suggests that benefits in industry could indeed be derived from studying quality management in the context of scientific data production, particularly in the area of standards research.

By exploring the nature of SDPs at the National Radio Astronomy Observatory in Green Bank, W. Va., examining the process by which they are produced, and comparing the findings with other scientific data production processes, the following fundamental assertions can be made:

- Scientific data and their production process are unique in three ways: unlike TDQM, the production process encompasses the generation of raw data and products derived from those data; the lifetime of the data product stored in an archive will be longer than the lifetime of the instruments that generated the data, and their quality must be evaluated subjectively as well as objectively; and the continuous improvement process must include the refinement of data products in an archive, and the integration of continuous learning into the production process to enable the refinement.
- Scientific data can be classified as device monitoring data, raw data, calibrated data, derived data, assimilated data, and model output data. If a dataset is intended for consumption by a researcher-consumer, it can be considered an SDP.
- Fitness for use is an objective process until the time the raw data are produced, which may include calibration, depending upon the level of quality that calibration provides. After this stage, evaluating fitness for use becomes relative to the researcher's specific scientific intent.
- Characterizing the quality of an SDP intended for use by a consumer requires an understanding of the scientific intent for which it is being used, and the process that produced the SDP. The quality of the SDP is a function of the purpose for which the product is intended.
- A semiotic framework for understanding data quality objectives in the context of the scientific decision process is more appropriate than using data quality objectives alone, because it reflects which quality objectives are most significant at each stage of the SDP production process.
- The temptation to focus on mass inspection of produced SDPs for characterization alone should be avoided; this tendency will become particularly challenging as SDP production is commoditized by the science data center management model. The more effective strategy is to design quality into the SDP production process, which requires feedback of organizational learning into the production process.
- Policies must be defined for who owns data quality for each level of SDP. It may be reasonable for the production facility to be responsible for raw and/or calibrated data only, and for the researcher to be responsible for the quality of derived data. The production facility may not be capable of assuring data quality to a particular level for all future experiments.

Planned future work will build on the foundations provided by this study. For example, an accurate characterization of what constitutes scientific data quality is fundamental to the classification and examination of quality costs with respect to scientific data. Upcoming work will address the establishment of a quality cost system to help NRAO Green Bank identify how to better meet its quality objectives. With an understanding of what constitutes an SDP and science data quality, the relationship between data quality and software quality can also be investigated. This is an important link because software is used extensively to collect, assimilate, and analyze data. For example, software that does not produce results that satisfy a researcher's intent cannot be considered high quality, even when it meets all requirements on the surface and has been produced by a high-performance team operating effectively and efficiently. Important links such as this have not yet been addressed in the literature.

ACKNOWLEDGMENTS

The author wishes to thank the following individuals for their review of this manuscript: Amy Shelton, deputy division head for software development, and Dr. Richard M. Prestage, assistant director, at the National Radio Astronomy Observatory in Green Bank, W.Va.; and Dr. Jim Braatz and Dr. Robert Garwood, scientists, National Radio Astronomy Observatory in Charlottesville, Va. The National Radio Astronomy Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. In addition, the efforts of Ron DuPlain (an undergraduate computer engineering co-op student from the University of Cincinnati) cannot go unrecognized. DuPlain's efforts and early returns analyzing telescope device interactions and preparing an automated quality diagnostics software package have been critical in the development of many of the ideas presented in this article.

REFERENCES

Anderson, D. M. 2002. Mass customization. In *Build-to-Order & Mass Customization: The Ultimate Supply Chain Management and Lean Manufacturing Strategy for Low-Cost On-Demand Production without Forecasts or Inventory.* Cambria, Calif.: CIM Press.

Becker, S. 1998. A practical perspective on data quality issues. *Journal of Database Management* 9, 35-37.

Cykana, P., and M. Stern. 1996. DOD guidelines on data quality management. In *Proceedings of the 1996 Conference on Information Quality,* Cambridge, Mass.

Dvir, R., and S. Evans. 1996. A TQM approach to the improvement of information quality. In *Proceedings of the 1996 Conference on Information Quality,* Cambridge, Mass.

English, L. 1999. *Improving data warehouse and business information quality: Methods for reducing costs and increasing profits.* New York: John Wiley & Sons, Inc.

English, L. 2004. Data: An unfolding quality disaster. *DM Review* (August). Available at: http://www.dmreview.com/article_sub.cfm?articleId=1007211.

Evans, J. R., and W. M. Lindsay. 2005. *The management and control of quality.* Mason, Ohio: South-Western College Publishing.

Graefe, G. 2003. Incredible information on the Internet: Biased information provision and a lack of credibility as a cause of insufficient information quality. In *Proceedings of the 8th International Conference on Information Quality (ICIQ-03),* Cambridge, Mass.

Gray, J., D. T. Liu, M. A. Nieto-Santisteban, A. S. Szalay, G. Heber, and D. DeWitt 2005. Scientific data management in the coming decade. *Microsoft Research,* MSR-TR-2005-10. Available at: http://research.microsoft.com/research/pubs/view.aspx?tr_id=860.

Hanuschik, R., A. Kaufer, A. Modigliani, S. D'Odorico, and H. Dekker 2002. Quality control of VLT-UVES data. In *Proceedings of SPIE 4844, Observatory Operations to Optimize Scientific Return III,* Waikoloa, Hawaii.

Institute for the Study of Society and Environment (ISSE). 2005. Societal Aspects of Weather Project Description. Available at: http://www.isse.ucar.edu/e_socasp.jsp.

Juran, J. M., and F. M. Gryna. 1988. *Juran's quality control handbook,* fourth edition. New York: McGraw-Hill.

Kim, W., and B. Choi. 2003. Towards quantifying data quality costs. *Journal of Object Technology* 2, no. 4 (July-August): 69-76.

Kujala, J., and P. Lillrank. 2004. Total quality management as a cultural phenomenon. *Quality Management Journal* 11, no. 4: 43-55.

Loshin, D. 2001. Enterprise knowledge management: The data quality approach. San Diego: Morgan Kaufmann.

Loshin, D. 2004. Knowledge integrity: Data standards and data models. *DM Review* (January). Available at: http://www.dmreview.com/article_sub.cfm?articleId=7931.

Massachusetts Institute of Technology (MIT) Total Data Quality Management (TDQM) Program. 2005. What is TDQM? Available at: http://web.mit.edu/tdqm/www/about.shtml.

National Science Foundation (NSF), in Budget of the United States Government. 2006. Available at: http://www.whitehouse.gov/omb/budget/fy2006/nsf.html.

O'Neil, K. 2005. Personal communication, October 6.

Price, R. J., and G. Shanks. 2005. Empirical refinement of a semiotic information quality framework. In *Proceedings of the 38th IEEE Hawaii International Conference on System Sciences,* Hawaii.

Redman, T. C. 1996. *Data quality for the information age.* Norwood, Mass.: Artech House, Inc.

Shankaranarayan, G., M. Ziad, and R. Y. Wang. 2003. Managing data quality in dynamic decision environments: An information product approach. *Journal of Database Management* 14, no. 4:14-32.

Spivak, S. M., and F. C. Brenner, eds. 2001. *Standardization essentials: Principles and practice.* New York: Marcel Dekker, Inc.

Tody, D. 2004. NRAO Observatory Model v0.1 (internal distribution only).

Tsien, P. Y. 2004. Data management: The quest for quality. Accenture White Paper, August 24.

Wang, R. Y. 1998. A product perspective on total data quality management. *Communications of the ACM* 41, no. 2: 58-65.

Wang, R. Y., and D. M. Strong. 1996. Beyond accuracy: What data quality means to data consumer. *Journal of Management Information Systems* 12: 5-34.

BIOGRAPHY

Nicole Radziwill is the division head for software development at the National Radio Astronomy Observatory (NRAO) in Green Bank, W.Va. Prior to NRAO, her experience includes managing consulting engagements in customer relationship management and sales force automation for telecommunications clients of Nortel Networks, and working in scientific computation at the National Oceanic and Atmospheric Administration (NOAA) Forecast Systems Laboratory in Boulder, Colo. She has more than a decade of experience managing continuous improvement efforts in business and technology, specializing in software development and process improvements that result from information technology. Radziwill has a degree in meteorology, an MBA, and is currently pursuing a doctorate in technology management and quality systems. She is a frequent speaker at conferences in astronomy and information technology, and is recognized as an ASQ Certified Quality Manager. She can be reached by e-mail at nradziwi@nrao.edu .

APPENDIX 1
TIQM in the context of one scientific data production process

	Process
Planning (P1 & P3): To understand the nature of the SDP production process; how quality issues arise, are detected, and are remedied at each stage, and how to improve the process to yield higher quality SDPs in the future.	**A.** Understand all factors that impact the SDP production process, in particular, the various causes of variation impacting the SDPs. This knowledge is required to be able to distinguish common causes of variation from assignable causes, and only launch improvement efforts for those apparent failures that are not due to normal and expected variation. • **Environmental factors.** Some are controllable over time, some are not (such as weather). Because people must adapt to the environmental factors they cannot control. • **Device health.** Are the components of the instrumentation performing to specification, and collecting raw data at the right accuracy, precision, and sampling frequency? Are they performing as expected over time? • **Nature of the instrument.** Know the capabilities and limitations of the devices producing the raw data, so one does not attempt to optimize beyond what has been designed. • **Processing stages.** Understand the assumptions behind algorithms used. Be aware of steps that might introduce numerical instability. Understand the sampling frequency required to capture features at each stage of the process, ensuring that desired features are resolved. • **Human causes of variation.** Understand the degrees of structure and flexibility in how the data are collected and SDPs are constructed. There is skill associated with operating an instrument or controlling a model so that scientific intent is realized, and investigators may need training. Training must be consistent so that those involved in the SDP production process do not contribute added variation because they followed different practices. • **Communication of quality information to consumer.** Determine a means to effectively characterize the quality of an SDP at each level in a way that helps the consumer subjectively establish the fitness for use for a particular SDP. This characterization will be unique for each step of each data production process, and represents a specialized research problem. **B.** Establish a process for identifying, documenting, and encouraging best practices for data quality improvement in each of the domains articulated above. Plan for a learning organization, in which all members of the data production organization can and will contribute to SDP quality improvement. In an environment where participation and teamwork are encouraged, the contributions from the scientist who uses the SDP in his or her decision process cannot be overlooked. A process must be in place to capture and act on the quality assessments that these individuals provide. **C.** Identify how to close the loop between the discovery of new SDP quality information and operations. This must involve feedback into the research process, in addition to operational policy development and software requirements generation. **D.** Establish a quality costs program to guide decision making in the assurance, control, and improvement stages, to ensure that optimal utility is achieved. **E.** Agree upon usage, ownership of data, and policies and processes that impact the production and application of data products. This requires an understanding of the dimensions of quality of information domains and information policy (outlined in Figure 1).
Assurance and Control (P2): To assess the quality of the production process and resultant SDPs. To distinguish between operational problems and changes in the SDP due to expected variation in the SDP production process, and respond appropriately.	**A.** Objectively assess the quality of data types as they are being produced. Segment the assessments between raw and derived data, assimilated data and model output, and use the dimensions of quality that are pertinent to the level of SDP in review as per the semiotic definitions in Figure 4. **B.** Automate the objective assessment. Use this to establish the thresholds at which the data production process can, and should, become dynamic and self-repairing. **C.** Provide a means for ensuring subjective quality control. The customer's perception of quality may, in part, depend on the availability of supplemental information that will help the scientist make subjective quality assessments. This means that in addition to characterizing the quality of a produced SDP, supplemental SDPs must be identified and made available.
Improvement (P5 and P6): To accelerate the delivery of high-quality SDPs to the consumer.	**A.** Establish and manage an effective process for identifying and responding to the right operational issue at the right time. **B.** Participation from everyone in the organization can be achieved by developing "rules-based" applications to immediately integrate subjectively determined quality information back into the production process. **C.** TIQM's 14 points of information quality should be applied to encourage continued development of an information quality culture.

APPENDIX 2
TIQM's 14 Points of Information Quality (IQ)

TIQM's 14 Points of Information Quality, based on Deming's 14 Points which revolutionized the Japanese manufacturing environment after the Second World War, can be downloaded in poster form from http://www.dmreview.com/editorial/dmreview/200309/200309_TIQMposter.pdf.

1. Create constancy of purpose for improvement of information product and service. Must solve problems of today *and* problems of tomorrow; the obligation to the knowledge worker (information customer) never ceases.
2. Adopt the new philosophy of quality-shared information as a tool for (business) performance excellence. "Reliable quality shared information reduces (business) costs." This means a transformation of information systems management and operations management.
3. Cease dependence on inspection to achieve information quality. Design quality into data and application design and to information production processes.
4. End the practice of: 1) developing applications on the basis of "on-time," "within budget" measures alone without a measure of quality, and 2) capturing data at the lowest cost. Develop common data, create programs, and invest in and develop trust in information producers.
5. Improve constantly and forever the processes of application and data development and service, and information production and maintenance. This will result in a continual reduction of costs of information scrap and rework, and will result in opportunity gain.
6. Institute training on information quality for all employees, especially management and information producers. If someone is to do a good job, they need to know how.
7. Institute leadership for information quality. Appoint a full-time information quality leader to lead the enterprise. Management must lead, not just supervise. Management must assume accountability for information quality.
8. Drive out fear of data uncertainty or data correction. Create a nonblame, nonjudgmental environment. Encourage risk to change without punishing missteps. Encourage improvement by eliminating root causes without blame.
9. Break down barriers between staff areas. Work as partners in teams, including information management and application development, IT, and business. Enterprise failures occur when departments work autonomously toward their own goals that suboptimize downstream process performance.
10. Eliminate slogans and exhortations and replace with actions for information quality improvement. Most information quality problems are caused by the system rewarding for speed, lack of clear data definition, lack of resources, training and understanding of downstream information customers, and quality requirements. Exhortations without management action for quality create stress and frustration with only temporary benefits. Provide commitment and resources for quality.
11. Eliminate quotas of "productivity" for information producers and management that increase errors and cost of information scrap and rework. Quotas for quantity actually decrease real productivity by decreasing quality that drives up the cost of scrap and rework. Create a balanced scorecard that includes end-customer satisfaction, internal information producer and knowledge worker satisfaction, and information scrap and rework elimination along with financial performance.
12. Remove barriers to pride of workmanship. Empower information producers to fix the problems in the processes, for they know the problems in their processes, and if given the opportunity will fix them. Develop a habit of information "defect prevention" as part of everyone's job.
13. Encourage vigorous education and self-improvement for all knowledge workers. Everyone must understand the paradigm shift and learn tomorrow's skills. Provide education in information age principles, value chain management, self-improvement, and process improvement.
14. Take action to accomplish the transformation for information quality. Senior management must organize itself for information quality to happen; senior management must feel the pain of the status quo; senior management must communicate to people why change is necessary. Implement a plan-do-check-act process for information quality improvement, and recognize that every process is a candidate for improvement.