NISS

Workshop Report: Affiliates Workshop on Data Quality

Alan F. Karr, Ashish P. Sanil, Jerome Sacks,
and Ahmed Elmagarmid

Technical Report Number 117
March, 2001

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

National Institute of Statistical Sciences www.niss.org

PO Box 14006 Research Triangle Park, NC 27709–4006 Tel 919.685.9300
FAX 919.685.9310

WORKSHOP REPORT

Affiliates Workshop on Data Quality

November 30 – December 1, 2000
Morristown, NJ

Alan F. Karr, Ashish P. Sanil and Jerome Sacks
National Institute of Statistical Sciences
{karr, ashish, sacks}@niss.org

Ahmed Elmagarmid
Purdue University
ake@cs.purdue.edu

March 15, 2001

Contents

1 Executive Summary

1.1 Workshop Purpose

The purpose of the workshop was to bring to the fore current issues of data quality as they affect users and researchers dealing with large data sets in potentially complex settings. There are (at least) three overlapping constituencies that have a stake in these matters: computer scientists, statisticians, and users (especially in business and government). One aim was to create interactions across the constituent groups leading to collaborative efforts to resolve open questions; another goal was to disseminate more widely an awareness of the problems, methodologies and practices.

1.2 Particulars

The Workshop was held at Telcordia Technologies' corporate center in Morristown, NJ, on November 30 and December 1, 2000. It was sponsored by the NISS affiliates program, with additional funds from NSF grant DMS–9904164 and Telcordia.

Members of the organizing committee, which was chaired by Ahmed Elmagarmid (Purdue) and Jerome Sacks (NISS), are listed in Appendix B.

The 46 attendees, from 26 organizations, of which 15 are NISS affiliates, included statisticians, computer scientists and owners of data quality problems. They are listed in Appendix C.

The complete workshop program appears in Appendix A; presentations are summarized in §2. Two tutorials were presented, by Richard Wang of Boston University (§2.1.1) and Larry P. English of Information Impact International, Inc. (§2.1.7).

Other formal presentations were made by Munir Cochinwala (Telcordia Technologies), William E. Winkler (US Census Bureau), Allan R. Wilks (AT&T Labs — Research), Ann Thornton (Deloitte & Touche LLP), Dean H. Judson (US Census Bureau), Vassilis Verykios (Drexel University) and Yannis Vassiliou (National Technical University of Athens).

An "open mike" session (see §2.2) enabled other participants to place their views of data quality before the group as a whole.

1.3 Findings

The principal findings of the workshop, elaborated in §3, are that:

1. *Data are a product*, with customers, to whom they have both cost and value.

2. As a product, data have *quality*, resulting from the *processes* by which data are generated.

3. *Data quality depends on multiple factors*, including (at least) the purpose for which the data are used, the user and time. Whether this implies that objective measures of data quality do not exist is unclear.

4. *Data quality is multi-dimensional*, and includes attributes of intrinsic quality, accessibility, context (For example, are the data relevant?) and representation (Are the data interpretable?).

5. In principle, *data quality can be measured and improved*, but the abstractions, techniques and software tools to do so are lacking.

6. *Human issues are central*, because people are the key links in the processes that generate most data.

7. Realization of importance of data quality is *growing but inconsistent*. Many organizations are aware that their data may be of less-than-desired quality, but few can characterize the resultant impact, and few have implemented programs to improve data quality.

1.4 Recommendations

Recommendations arising from the workshop are summarized here.

First (see §4.1 for details), NISS (or others) should conduct *additional workshops focused on case studies*, in order to (1) Characterize individual data quality problems; (2) Identify overarching features common to multiple data quality problems; and (3) Continue the development of the requisite, cross-disciplinary collaborations, and create links between the "owners" of data quality problems — industry and government — and academia, engaging additional sources of talent to attack them.

Second (see §4.2 for details), a broad but concrete *program of research* should be undertaken, addressing details of specific data quality problems, but with the goal of developing widely applicable tools. Fundamental issues to be addressed include:

- Definitions and abstractions for data quality that reflect the multi-scale nature of data quality and the multiple uses of information derived from the data; database formalisms and schema that incorporate data quality; and metadata abstractions that accommodate data quality.
- Quantification and measurement of data quality, by means of metrics for data quality that capture quality at multiple scales and represent the impact (for example, in economic terms) of data quality. From these metrics, specifications for data can be constructed that reflect multiple demands on data. For example, specifications should be able to distinguish administrative uses of data from inferential uses.
- Models for data quality that reflect the nature of the processes by which data are generated, and in particular the fundamental role of people in these processes. Corresponding statistical procedures are needed, for example, to estimate or verify data quality, or to evaluate the impact of data quality, as well as perform inference in the presence of low quality data.
- Models and algorithms to characterize changes in data quality under such transformations as aggregation and integration of multiple databases.
- Strategies for improvement of data quality, exploiting the perspective of data as product. An example is to identify and characterize controllable and uncontrollable sources of variability in data quality.
- Software tools that solve real data quality problems. Such tools (even in prototype form) can also be used to evaluate and improve new theory and methodology. Issues include algorithms

that cope with complexity of the techniques as well as the scale of the data and problems of human–computer interaction such as presentation and visualization.

Finally, *connections between data quality and other problems* should be explored.

2 Program

In this section, we present abstracts and summaries of the presentations at the workshop, which are available at www.niss.org/affiliates/dqworkshop/presentations.html. We also summarize key points from the December 1 "open mike" session, at which other participants highlighted their data quality concerns and problems.

2.1 Presentations

These appear here in the order in which they took place (see §A).

2.1.1 Richard Wang: Raising the Bar for Data Quality in the New Millenium

Abstract. The tutorial will provide an overview of the research conducted at the MIT Total Data Quality Management (TDQM) program over the last decade, emphasizing the management of information as a product. We will discuss the approach of the MIT TDQM program advocating the institutionalizing of TDQM programs for long-term benefits. Concepts such as multi-dimensional data quality, data quality metrics, evaluation of the user's assessment of data quality, and data production maps will be presented in this context. Research directions stemming from recent work such as Huang, Lee & Wang (1999), Wang, Ziad & Lee (2000), *Journey to Data Quality: A Roadmap to Higher Productivity* (in preparation), and *Data Quality in the Health Care Industry* (in preparation) will be touched on.

Summary. With striking similarity to the talk by English (§2.1.7), this presentation emphasized the necessity to *manage data as a product*, and stipulated that the data generation–to–consumption process should be managed in the same way as the process that produces any other product. For example, TQM concepts like the Deming cycle and statistical quality control can be applied to the data production process.

The talk also made clear that data quality not only goes far beyond conventional views of record-level accuracy, but also is multi-dimensional. Proposed as attributes were *intrinsic quality* (accuracy, objectivity, believability), *accessibility* (access, security), *contextual quality* (relevance, value added, timeliness, completeness) and *representational quality* (interpretability, consistency, manipulability).

A Total Data Quality Management (TDQM) framework was presented, based on a cycle of "Define, measure, analyze, improve."

Other ideas raised included: (1) Data as assets; (2) Specifications (e.g., quality) for information products; and (3) Extending the standard entity–relationship model of database design to incorporate data quality.

2.1.2 Munir Cochinwala: Data Quality and Reconciliation

Abstract. In this talk we will give an overview of the data quality research program at Telcordia. At Telcordia we use an in-house research prototype to handle data reconciliation and data quality analysis. The prototype and accompanying methodology include rapid generation of appropriate pre-processing and matching rules. The prototype uses a modular JavaBeans-based architecture that allows for customized matching functions and iterative runs that build upon previously learned information. Telcordia has been able to provide significant insights to clients who recognize that they have data reconciliation problems but cannot determine root causes effectively when using currently available off-the-shelf tools.

Summary. The framing problems for the presentation were (1) Linking customer databases of recently merged companies; (2) Linking magazine retailer and wholesaler databases of sales records (in order to identify gaps in reporting); (3) Reconciliation of amount of network usage and charges billed for usage. Their common feature is that databases are to be merged, and in the process, duplicates must be detected and errors removed.

An abstract process model for a generic data reconciliation / record linkage task was described, together with a framework of JavaBeans components to implement specific version of the task. The components include data sources, data filters, record-matching rules, classification rules and output data sets that implement the process model. Extensibility of the components was stressed: users can incorporate their domain-specific logic through JavaBean components implementing a given interface.

Other issues touched on included metrics for data quality (see §4.2.2), GIS databases, stream data, SQL-like functionality, and gaps between data quality for the system designer and for end users. For example, how can end users be educated about the meaning of notions of closeness used for record-matching?

2.1.3 William E. Winkler: Record Linkage Methods

Abstract. Record Linkage is used for identifying duplicates within files and merging sets of files. This talk describes methods and software. Names, addresses, and other components in a file are initially parsed into corresponding components such as first names and house numbers that are more easily compared. The model of Fellegi & Sunter (1969) generalizes recent work on Bayesian networks. It is used to get matching or classification scores that rank the relative quality of matches. In some situations, a generalized EM algorithm can be used to obtain optimal matching parameters through unsupervised learning. In other situations relatively small amounts of training data can be combined with large amounts of unlabelled data. String comparators account for partial agreement between strings. A generalized assignment algorithm optimizes sets of matches.

Summary. The talk proceeded from the fundamental matching model of Fellegi & Sunter (1969): given files **A** and **B** to be matched, pairs $x \in \mathbf{A}$ and $y \in \mathbf{B}$ are classified as matches M or non-matches U on the basis of the likelihood ratio

$$R = \frac{P(\gamma(x, y) \in \Gamma | M)}{P(\gamma(x, y) \in \Gamma | U)},$$

where γ is an agreement pattern and Γ a specified decision region, and where P represents a model for the data pairs. The optimal rule (Fellegi & Sunter, 1969) specifies thresholds T_* and T^*, with a match declared if $R > T^*$, a non-match declared if $R < T_*$ and clerical review required if $T_* < R < T^*$.

The case that agreement between different components of the data is conditionally independent given M or U is particularly amenable to efficient software implementation of Bayesian classifiers.

Within this setting, a variety of other implementation issues was addressed: (1) Parsing and standardization of names and addresses; (2) Address matching; (3) Accounting for specific errors (Example: typographical errors such as deletions, insertions and transpositions); (4) The need for "truth" data to train classifiers; and (5) Estimation of error rates in the data; and the need for efficient matching algorithms for large databases.

Issues raised in discussion included: (1) Whether and how to perform database-level matches instead of individual record-level matches; (2) Decision-theoretic approaches that would accommodate consequences of mismatch other than the probability of mismatch; and (3) Use of context-sensitive methods in matching algorithms, rather than only to prepare and standardize data.

2.1.4 Allan R. Wilks: Data Quality for Large Transaction Streams

Abstract. Data analysis for very large data sets is particularly challenging when the arrival of the data is relentless — it's like drinking from a fire hose. An example is the 50 GB/day stream of transaction call detail from the AT&T long distance network. This talk will describe data quality issues for this stream from several perspectives, including: timely detection of data anomalies; maintaining data integrity through software and hardware integrity; and the impact of subject-matter expertise on quality of results.

Summary. The data represent more than 300,000,000 transactions per day, collected by more than 400 telephone switches, and are in a complicated, variable-record-length format. The problem was to identify and explain *all* anomalies (especially gaps) in the data stream (for example, for use in fraud detection).

Beyond scale, complications include different reporting frequencies for different sources, the complex data format and the difference between database access and streaming access to data.

Multiple tools are employed: streaming tools, such as shell scripts, `awk` and C programs; the Daytona database management system; and alerting methods, such as software failure alarms and pagers, for automating the gap detection process.

With these tools, users of the data were able to detect small gaps in the data very often. Other essential needs were: (1) To bring users' domain knowledge to bear on the problem; (2) The importance of regular and systematic user feedback; and (3) Keeping all raw data and logs, in order to trace the source of data gaps and to explain anomalies.

2.1.5 Ann Thornton: Challenges in Improving Information Quality

Abstract. While most agree that good data is preferable to bad data, an organization that sets out to assess and/or improve information quality faces several challenges:

- Defining the importance of information quality: Assessing the costs and benefits of improvements to information quality is difficult and requires participation of those using the information as well as those contributing information.
- Assessing information quality: A daunting amount of analysis can be required to perform a detailed assessment (for example, at the data element level). Therefore, assessments of the importance of each data element are used to prioritize work effort.
- Addressing information quality problems: Performing root cause analysis and appropriate corrective actions is often preferable to short-term fixes, but process improvement requires commitment.
- Ongoing measurement and monitoring: Metrics for information quality can be difficult to construct.

Summary. This talk was an enterprise-wide view of data and information quality (see also Loshin (2001)), focussing on implementation issues. It highlighted the difficulties of addressing data quality issues related to implementations of corporate enterprise resource planning (ERP) and customer relationship management (CRM) systems. Information/data quality task forces were discussed as one mechanism for helping organizations deal with data quality issues.

Other issues raised were: (1) How to measure the costs and benefits of information quality; (2) The importance of simple ideas (examples: data range and type checks); (3) Tradeoffs between repairing bad data and "demanding, facilitating, and researching better data collection;" (4) The role of domain knowledge (e.g., to detect outliers); and (5) Visualization as a tool to understand and evaluate data quality.

Also presented was a multi-resolution perspective on data quality: when data are aggregated or summarized (for example, to produce management reports), what are the quality implications? In particular, less-than-perfect data may suffice in such circumstances.

The importance of high-quality metadata and data dictionaries ("metadata quality") emerged clearly in discussion.

2.1.6 Dean H. Judson: The Statistical Administrative Records System and Administrative Records 2000 Experiment — System Design, Successes and Challenges

Abstract. The Statistical Administrative Records System (StARS) is an attempt to link six major federal administrative records files together into a composite database that will be used to simulate an administrative records census. The six major files include the IRS 1040 master file, the IRS 1099 and information returns file, the Medicare beneficiary database, the Selective Service System registrant file, the HUD–TRACS tenant rental assistance file, and the Indian Health Service file. A seventh file, the Census NUMIDENT (which is a translation of the Social Security Administration NUMIDENT file), is used as a "lookup" file for social security number (SSN) validation and as a source of demographic characteristics.

These six files are edited to create standardized person and address records, unduplicated, validated, missing fields are imputed, and finally merged to create statistical composite records. The

presentation will focus on the challenges that each of these processes create, including thresholds for observation, database ontologies, unduplication and matching decisions, and the relationship between a dynamic database and a dynamic population. The presentation will conclude with a brief description of the Administrative Records Experiment (AREX 2000), which is a test of an administrative records census in two test sites in 2000.

Summary. A substantial part of the presentation dealt with difficulties encountered in merging the six databases, as well as the steps taken to overcome the problems. The major problems included: (1) Multiple ways of recording legitimate addresses in the different databases; (2) Minor variations in records (e.g., spelling) that make linkage difficult; (3) "Pure" inconsistencies (e.g., multiple sexes recorded for the same individual); and (4) The dynamic nature of the databases. (example: addresses change frequently.)

Discussion revolved around how to use these lessons, and how abstract the general structure of this problem in order to address the general data merging problem.

2.1.7 Larry P. English: Information Quality Processes and Technologies — Information Quality in Practice

Abstract. It is clear that information quality is no longer irrelevant nor a luxury in information systems. It is a requirement for sustainable competitive advantage in the new economy. The Industrial Age matured when quality processes such as continuous process improvement (CPI), total quality management (TQM), and Kaizen transformed manufacturing processes, eliminating the costs of scrap and rework. We are now seeing the maturing of the Information Age as a result of applying the same quality processes to the information product — the new currency of the new economy.

This presentation presents the state-of-the-art in information quality improvement processes as applied by leading edge companies. Processes for information quality assessment and improvement are described, as are the classifications of information quality technologies, describing the strengths and limitations in information quality management. Organizations around the world who have implemented successful information quality processes are used to illustrate techniques and cultural transformation required for a sustainable information quality environment:

- Information quality: what it is and what it isn't;
- Trends in information quality processes and methods;
- Classifications of and trends in information quality technologies;
- Information quality improvement: the maturing of information management;
- Systemic culture change requirements to sustain information quality initiatives.

Summary. The talk presented a detailed view of what constitutes information quality: "quality in all characteristics of information, such as completeness, accuracy, timeliness, clarity of presentation" in a manner that "consistently meets knowledge worker and end-customer expectations."

Central to this view is the concept of data and information as products, which points to a plethora of issues and approaches, such as: (1) The lifecycle of data from data generation to data

customer as a production process; (2) The application of TQM to data production processes, such as an emphasis on creating data quality at the start, rather than by re-work (the principal means currently); (3) Measuring the cost of non-quality data; (4) Human factors aspects of data quality process; and (5) Accountability (even warranties) for information.

In response to a question, a distinction between data and information was proposed:

$$\text{information} = \text{data} + \text{context}.$$

2.1.8 Vassilis Verykios: A Decision Model for Cost Optimal Record Matching

Abstract. In an error-free system with perfectly clean data, the construction of a global view of the data consists of linking or joining two or more tables on their key fields. Unfortunately, most of the time, data stored in real life database systems are neither carefully controlled for quality nor necessarily defined commonly across different data sources. As a result, the creation of such a global data view resorts to approximate joins. In this talk, an optimal solution is proposed for the matching or the linking of database record pairs in the presence of inconsistencies, errors, or missing values in the data.

Existing models for record matching rely on decision rules that minimize the probability of error, which is the probability that a pair of records is assigned to the wrong class. Often in practice, minimizing the probability of error is not the best criterion to design a decision rule because the misclassifications of different samples may have different consequences. In this talk, we present a decision model, which minimizes the mean cost of making a decision, by assigning a different cost to each kind of misclassification. More specifically, (a) we present a decision rule, (b) we prove that this rule is optimal with respect to the cost of the decision making process, and (c) we compute the probabilities of the two types of errors that occur when this rule is applied. We also present a closed form decision model for a class of comparison vectors having conditionally independent binary components and finally we demonstrate some experimental results from applying the proposed model.

Summary. The fundamental tool used is a *comparison vector*. The comparison vector for records $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ from databases with the same schema is

$$\mathcal{C}(x, y)_i = \begin{cases} 1 & \text{if } x_i = y_i \\ 2 & \text{if } x_i \neq y_i \\ 3 & \text{if } x_i \text{ or } y_i \text{ is missing.} \end{cases}$$

If X and Y are random (drawn from some population) then so is $\mathcal{C}(X, Y)$. A Bayes classifier was constructed, based on the likelihood ratio approach of Fellegi & Sunter (1969), for the case that $\mathcal{C}(X, Y)$ has one distribution when X and Y "match" and another when they do not, and where costs are associated with various incorrect decisions.

2.1.9 Yannis Vassiliou: Developing Data Warehouses with Quality in Mind

Abstract. Data warehouses provide large-scale caches of historic data. They lie between information sources gained externally or through online transaction processing systems (OLTP), and

decision support or data mining queries following the vision of online analytic processing (OLAP). In developing and operating data warehouses, one can distinguish between different processes that raise quality considerations. Using the framework of a general formal architecture developed in the DWQ project, this talk discusses some of the research conclusions regarding development of data warehouses considering also quality factors. Basic processes and the related quality dimensions are considered. The quality factors, metrics and measurement methods are also presented.

Summary. The talk presented work done on the Data Warehouse Quality (DWQ) project by researchers based at seven European universities and research organizations, describing the frameworks and tools developed and listing the research conclusions.

The framework stipulates that quality, processes and system architecture be considered as three critical components ("metamodels") of data warehouse design. For each component, guidelines were presented for design, which were based on classifying the subcomponents by location (source, enterprise, client) and nature (physical, logical, conceptual).

Other points emphasized were (1) The interplay between quality factors and design/evolution characteristics of the warehouse; (2) The extreme importance of effective metadata management facilities; (3) Quality models broad to encompass data, processes and service (to users); (4) Potential tension between "scientific" (system-designer-created) definitions of data quality and those of other stakeholders; and (5) Quality for metadata.

The need was emphasized for quality metrics encapsulating such factors are correctness, minimality, traceability, accessibility, availability, security, usefulness and interpretability.

2.2 Open Mike Session

Linda Nelsen described a number of data quality issues in the pharmaceutical industry. A principal goal is to match multiple databases (from the manufacturers, clinical and pre-clinical trial data, marketing, manufacturing and outcome research data; from external sources such as insurance companies, plan and billing information available without test results, diagnoses, or actual prescribed use of drug) in order to identify diagnoses and actual drug usage. Data quality problems are rampant, ranging from incorrect data entry to coding and billing biases, inconsistencies and missing values.

Suzanne Markel–Fox called attention to a number of reports and studies (GAO, 2000a,b,c) of the US General Accounting Office (GAO) on data quality and standards within the Federal government.

Focusing on medical data, Fox amplified Nelsen's concerns about the disconnect between claims data and clinical data, which may have serious health implications.

Domenico Natale discussed work on public administration services — specifically, tax information systems in Italy. The central task was to keep databases complete, accurate and current. Data quality problems arose from variety (some pay taxes directly, others indirectly), scale (200,000,000 paper documents and 30 fiscal domains) and timeliness (address changes occur for 10% of records per year).

David Banks described how the Bureau of Transportation Statistics (BTS) is evaluating 60 De-

partment of Transportation (DOT) databases. To provide a quick initial assessment, each database was scored on a 3-point scale according to criteria of (1) Timeliness; (2) Statistical collection methods; (3) Quality of documentation; (4) Accessibility; (5) Relevance to various mandates of the DOT; (6) Whether there is periodic structural review; (7) Validation procedures employed for the data; and (8) Cost.

Banks also noted the existence of robust statistical procedures that can provide informative inferences even in the presence of high error rates in the data (at least under certain assumptions).

Mohamed Elfeky presented a conceptual design for data quality management systems (DQMS).

3 Findings

Principal findings of the workshop are presented in this section.

1. *Data are a product*, and as such should be viewed and treated in the same way as any other product. In particular, data have (not always readily) identifiable *customers*, who use and pay for — even if indirectly, as for Federal databases — the data; *cost* to the customers, whether data are generated directly or purchased from another source; and *value* to the customers, which may be less consonant with cost than for many other products.

   This message was explicit in the tutorials by Wang and English, and implicit in many other presentations.

2. As a product, data have *quality*, in the same way as other products do. Such quality results from the *processes* by which data are generated.

   Wang and English delivered this message explicitly, while other presentations (Winkler, Judson and Nelsen, Fox–Markel and Banks in the open mike session) stressed the unusual nature of these processes, for example, the central role of people.

3. Data quality can, in principle, be *measured and improved*, but currently lacks a rigorous definition. Existing techniques and strategies (example: characterization and control of variability in the processes that generate data) for quality improvement are applicable in principle to data, but have rarely been applied.

   The importance of and need for metrics of data quality was perhaps the most universal message from the presentations, appearing in some form in all. Cochinwala, Winkler and Wilks described particular attempts to ameliorate data quality problems. Vassiliou alone described a system that attempts *ab initio* to deal with data quality.

4. Realization of importance of data quality is *growing but inconsistent*. Awareness of data quality surpasses both ability to characterize its impact and availability of methods to address it.

   Thornton, in particular, described a setting of acute awareness but limited capability to act.

5. In universities, recognition of data quality as an important area of research is virtually non-existent. In part, this results from "ownership" and recognition of data quality problems residing largely in industry and government.

6. *Data quality is contextual:* the same data may be adequate for one purpose or customer or at one time, but not of sufficiently high quality for other purposes or other customers, or at other times. In particular, there is a clear distinction between data quality for *administrative* purposes, such as customer billing, and data quality for *inference* purposes, such as analysis of buying patterns.

   This message was virtually universal; Cochinwala, Winkler and Wilks raised it explicitly.

7. *Data quality is multi-dimensional*, encompassing such attributes as accuracy, credibility, accessibility, relevance, timeliness and interpretability.

   Wang and Banks (open mike session) delivered this message explicitly. The diversity of interpretations of data quality among other participants adds further confirmation.

8. *Data quality is multi-scale*, ranging from individual records to databases to beyond, when multiple databases are integrated.

   This point perhaps emerged more strongly in discussions than in presentations: without being able (yet) to say what it is, there is clear consensus that data quality is more than record-level accuracy, and that more global (aggregated) concepts and metrics must be developed.

9. *Human issues are central*. More so than many products, data are generated by processes in which humans are the key links. Examples range from surveys to medical data entered by hospital personnel. In particular, there is a clear distinction between human-collected data and machine-generated data.

   English was insistent about this point; Nelsen and Markel–Fox confirmed it strongly.

4 Recommendations

The workshop made clear that the study of data quality is strongly cross-disciplinary, involving the fields of statistics, computer science (databases, artificial intelligence and human–computer interaction, in particular), in addition (for any specific problem) to domain knowledge and customers. Moreover, in light of the current inchoate state of the subject, *research on data quality must proceed up from concrete instances* before proceeding down from abstractions and theory.

The recommendations below reflect these points.

4.1 Further Workshops

This workshop succeeded at socialization and established initial lines of communication. These steps are a necessary beginning, given the diversity of views (even within each of the communities represented) of the problem of data quality.

The workshop did not, however, create a definitive research agenda (although some components of such an agenda are described in §4.2). To help advance the subject, *additional workshops focused on case studies* seem essential.

By focusing on case studies, these workshops will: (1) Characterize individual data quality problems in sufficient detail that initial research steps can be taken; (2) Identify over-arching features common to multiple data quality problems, without which the field will remain a collection of special cases; and (3) Continue the development of the requisite, cross-disciplinary collaborations, and at the same time, create links between industry and government, where data quality problems reside, and academia, where significant untapped talent exists to attack them.

4.2 Research Needs

Here we describe research issues identified at the workshop as meriting significant cross-disciplinary attention. Implicit here is the premise that data quality is not simply a set of similar but ultimately disconnected special cases; whether this will turn out to be true is not (yet) known.

4.2.1 Foundation Issues

The abstractions, concepts and definitions necessary to deal with current, let along future, issues of data quality are lacking.

Abstractions and definitions for data quality are required with multiple characteristics. First, it is necessary to distinguish among data quality at the record level, data quality at the database level and data quality at the level of integrated databases. Seconds, definitions must reflect the use(s) of information derived from the data (and hence also the users, time . . .), as well as accommodate domain knowledge properly. The abstractions must also support construction of implementable metrics for not only data quality itself but also its impact (§4.2.2).

Database formalisms and schema must be constructed that incorporate data quality. Record-level quality could, for example, be added as an attribute in relational tables, but how to treat more global measures is unclear.

Metadata abstractions must also accommodate data quality. For some geographical databases, data quality is represented in rudimentary form in the metadata (Dublin Core, 2001), but no general approach exists.

An intriguing question is whether existing and emerging formalisms for model evaluation and validation have analogues for data: is there such a thing as data validation?

4.2.2 Quantification and Measurement

Quantification and measurement of data quality are also strongly cross-disciplinary issues.

Metrics for data quality are necessary that: (1) Capture quality at multiple scales (see §4.2.1) and in particular go beyond record-level accuracy; and (2) Represent the *impact of data quality*, in either economic or other terms.

Specifications for data, constructed from the metrics, must reflect adequately multiple demands on data. For example, specifications should distinguish administrative uses of data from inferential uses.

4.2.3 Models of Data Quality

These are the most pressing and challenging gap.

Models for data quality must be developed, that reflect the nature of the processes by which data are generated. For example, the central role of people in many data generation processes (§3) must be accommodated. In a different direction, typical models for measurement error do not seem well suited to data quality problems: for example, what error model represents the transposition of digits by a data entry clerk, or the improper linkage of records? More generally, conventional statistical concepts of "noise" in data may not be rich enough to address some data quality problems.

Models for changes in data quality under transformations of the data are also essential. Examples of such transformations include: (1) Aggregation, which from a naive statistical perspective might be assumed to improve quality, but there seems to be no logical reason why this should be so; (2) Joins of relational tables within the same database; and (3) Integration of multiple databases, which also seems to be believed *a priori* to improve quality, but again this belief seems to reflect hopefulness more than necessity.

Inference procedures that accompany these models are needed, in order to deal with data of either characterized or unknown quality. These might build as well on existing techniques from robust statistics.

4.2.4 Improvement of Data Quality

Although to some degree this is an over-simplification, currently record linkage across multiple databases is the principal strategy for improving (record-level) data quality. From the perspective of data as product (§3), many more existing strategies (or modifications of them) for quality improvement can be brought to bear.

Models described in §4.2.3 are the basis for this proactive approach to data quality. This approach would, for example, emphasize prevention of problems — for example, through more informed data collectors, as opposed to the reactive approach of re-working existing data to ameliorate quality problems.

Variability — from both controllable and uncontrollable sources that affect data quality — needs to be identified and characterized.

4.2.5 Implementation

Ultimate impact (on customers) of *data quality strategies* — abstractions, quantifications, models and improvement methods — requires that they be implemented as software tools that can be applied to real problems. These tools (even in prototype form) are also essential to evaluation and refinement of data quality strategies. Associated research issues include:

Algorithms must be built that can cope with complexity of the techniques and procedures as well as the scale of the data.

Tools for presentation and visualization are necessary because the customers for data as product are diverse and often not sophisticated technically. Among the issues is invention of effective visual metaphors for data quality.

In addition to other disciplines noted already, significant input will be needed from collaborators having expertise in human–computer interaction, numerical computation and computer graphics/visualization.

4.2.6 Connections with Other Problems

Despite numerous unique characteristics, data quality is connected (possibly unexpectedly) with at least two other important classes of problems. Exploring these and other connections provides a way to leverage any research on data quality.

Data confidentiality is a problem of long standing to government agencies, and is burgeoning in the world of E-commerce. Here a central issue is to allow informative inference from confidential data without compromising confidentiality of the data (NISS, 2001).

The relationship between data confidentiality and data quality is complementary: the same tools that increase quality threaten confidentiality. For example, releases of even altered microdata (individual records) threaten confidentiality because these records can be re-identified by linking them to another database. Tools that alter data but allow informative inference (swapping of some attributes between records, for instance) point to ways of characterizing how much information can be extracted from low quality data.

Software quality bears strong resemblance to data quality. Both products are electronic rather than physical, so that quality is a characteristic of a class rather than instances. For both, quality is highly situational: in the same way that the quality of one database may be adequate for one purpose but not another, the same software may serve adequately in one setting, but be inadequate in another. For both products, the human element (data collectors and software developers) in the production process is both central and an essential source of variability.

In particular, the protracted (and still incomplete) effort to produce metrics and standards for software quality (Kan, 1994) may provide important insight into metrics for data quality.

Scientific misconduct in which data are falsified or altered, is a problem with significant economic and health consequencesMarshall (2000). Low quality and "fabricated" data may have similar characteristics, and the measurement of the one may provide means of detecting the other.

Acknowledgements

Staff members at Telcordia (especially Anne Marie Smith) and NISS (Katherine Kantner) made essential contributions to the smooth planning and execution of the workshop.

Partial financial support for the workshop was provided by the National Science Foundation, through grant DMS–9904164 to NISS, and by the NISS affiliates program.

References

Davis, J. R., Nolan, V. P., Woodcock, J. and Estabrook, R. W., eds. (1999). *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report*. National Academy Press, Washington.

Dublin Core Metadata Initiative (2001). Information available on-line at www.purl.org/dc/.

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

English, L. P. (2000). Seven deadly misconceptions about information quality. Available on-line at www.niss.org/affiliates/dqworkshop/papers.html.

Fellegi, I. P., and Sunter, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

US General Accounting Office (2000a). *Benefit and Loan Programs: Improved Data Sharing Could Enhance Program Integrity*. Report HEHS-00-119. Available on-line at www.gao.gov.

US General Accounting Office (2000b). US Customs Service: OR&R Needs to Resolve Timeliness and Data Problems Involving Headquarters Rulings. Report GGD-00-181. Available on-line at www.gao.gov.

US General Accounting Office (2000c). National Practitioner Data Bank: Major Improvements Are Needed to Enhance Data Bank's Reliability. Report GAO-01-130. Available on-line at www.gao.gov.

Huang, K.-T., Lee, Y. W., and Wang, R. Y. (1999). *Quality Information and Knowledge Management*. Prentice Hall, Upper Saddle River, NJ.

Judson, D. H. (2000). The statistical administrative records system: system design, successes and challenges. Available on-line at www.niss.org/affiliates/dqworkshop/papers.html.

Kan, Stephen H. (1994). *Metrics and Models in Software Quality Engineering*. Addison–Wesley, Reading, MA.

Liepins, G. E., and Uppuluri, V., eds. (1991). *Data Quality Control: Theory and Pragmatics*. Dekker, New York.

Loshin, D. (2001). *Enterprise Knowledge Management*. Morgan Kaufmann, San Diego.

Marhsall, E. (2000). How prevalent is fraud? That's a million–dollar question. *Science* **290** 1662–1663.

National Institute of Statistical Sciences (2001). Digital Government project Web site. Available on-line at www.niss.org/dg.

Redman, T. C. (1997). *Data Quality for the Information Age*. Artech House, Norwood, MA.

Redman, T. C. (2000). *Data Quality: A Field Guide*. Digital Press, Boston.

Wang, R. Y., Ziad, M., and Lee, Y. W. (2000). *Data Quality*. Kluwer Academic Publishers, Amsterdam.

Winkler, W. E. (2000). Machine learning, information retrieval and record linkage. Available online at www.niss.org/affiliates/dqworkshop/papers.html.

Appendices

A Workshop Program

Thursday, November 30, 2000

8:30 AM	Welcome and Opening Remarks Jon Kettenring, Telcordia, and Jerome Sacks, NISS
9:00	Tutorial I: Raising the Bar for Data Quality in the New Millenium Richard Wang, Boston University
10:45	Break
11:15	Data Quality and Reconciliation Munir Cochinwala, Telcordia Technologies
12:15 PM	Lunch
1:15	Record Linkage Methods William E. Winkler, U.S. Census Bureau
2:15	Data Quality for Large Transaction Streams Allan R. Wilks, AT&T Labs — Research
3:15	Break
3:30	Challenges in Improving Information Quality Ann Thornton, Deloitte & Touche LLP
4:30	The Statistical Administrative Records System and Administrative Records Experiment 2000: System Design, Successes and Challenges Dean H. Judson, US Census Bureau
5:30	Adjourn
6:00	Reception and Workshop Dinner

Friday, December 1, 2000

8:30 AM	Tutorial II: Information Quality Processes and Technologies — IQ in Practice Larry P. English, Information Impact Inc.
10:00	Break
10:15	A Decision Model for Cost-Optimal Record Matching Vassilis Verykios, Drexel University
11:15	Developing Data Warehouses with Quality in Mind Yannis Vassiliou, National Technical University of Athens
12:15 PM	Working Lunch and “Open Mike” Session
1:45	Plans for Report and Follow-up Alan Karr, NISS
3:00	Adjourn

B Organizing Committee

Mary Ellen Bock (Statistics, Purdue University)
John Chambers (Bell Labs, Lucent Technologies)
Dean Judson (US Census Bureau)
Sid Dalal (Telcordia Technologies)
William Eddy (Statistics, Carnegie Mellon University)
Ahmed Elmagarmid, co-chair (Computer Science, Purdue University)
John Eltinge (Bureau of Labor Statistics)
Jon Kettenring (Telcordia Technologies)
Joseph McCloskey (National Security Agency)
Munir Cochinwala (Telcordia Technologies)
Linda Nelsen (Merck & Company)
Brent Pulsipher (Pacific Northwest National Laboratory)
Jerome Sacks, co-chair (NISS)
Dennis Shasha (Computer Science, New York University)
Richard Wang (Management Information Systems, Boston University)

C Participant List

Name	Organization
Paola Angeletti	Societá Generale d' Informatica (SOGIE)
David Banks	Bureau of Transportation Statistics (*)
Robert Bell	AT&T Labs — Research (*)
Andrew Borthwick	ChoiceMaker Technologies, Inc.
John Chambers	Bell Labs, Lucent Technologies (*)
Kevin Coakley	National Institute of Standards and Technology (*)
Munir Cochinwala	Telcordia Technologies (*)
Sid Dalal	Telcordia Technologies (*)
Lorraine Denby	Avaya Labs (*)
Adrian Dobra	Carnegie Mellon University (*)
Mohamed Elfeky	Purdue University
Ahmed Elmagarmid	Purdue University
John Eltinge	Bureau of Labor Statistics (*)
Larry English	Information Impact International, Inc.
Arthur Goldberg	New York University
Hui Huang	Duke University (*)
David James	Bell Labs, Lucent Technologies (*)
Dean Judson	Census Bureau (*)
Alan Karr	NISS
Marc Kennedy	NISS
Jon Kettenring	Telcordia Technologies (*)
Tim Krick	Deloitte & Touche LLP
Li Liu	NISS
David Loshin	Knowledge Integrity, Inc.
Suzanne Markel–Fox	SmithKline Beecham (*)
Joseph McCloskey	National Security Agency (*)
Paolo Missier	Telcordia Technologies (*)
Domenico Natale	Societá Generale d' Informatica (SOGIE)
Nagaraj Neerchal	University of Maryland Baltimore County (*)
Linda Nelsen	Merck & Company (*)

Organizations marked by (*) are NISS affiliates.

Participant List (continued)

Name	Organization
Brian Park	NISS
Jennifer Pittman	NISS
Brent Pulsipher	Pacific Northwest National Laboratory (*)
Robert Rodriguez	SAS Institute (*)
Jerome Sacks	NISS
Ashish Sanil	NISS
Dennis Shasha	New York University
Surendra Singh	Telcordia Technologies (*)
Ann Thornton	Deloitte & Touche LLP
Eric Tollar	Telcordia Technologies (*)
Yannis Vassiliou	National Technical University of Athens
Vassilis Verykios	Drexel University
Richard Wang	Boston University
Allan Wilks	AT&T Labs — Research (*)
William Winkler	Census Bureau (*)
Michael Zhu	Purdue University (*)

Organizations marked by (*) are NISS affiliates.

D Selected Web Sites

www.census.gov/srd/www/byyear.html: Census Bureau – papers by Winkler and others

www.dataquality.com: Data Quality news and journal

www.gao.gov: General Accounting Office reports

www.infoimpact.com/index_flash.html: Information Quality Conference 2001 (Baltimore, September 30 – October 4, 2001)

www.mhsip.org: Mental Health Statistics Improvement Program, with several items relating to data quality

cii-server5.nci.nih.gov:8080/cde_browser/cde_java.show: National Cancer Institute Common Data Elements (CDE) Dictionary

www.statcan.ca/english/conferences/symposium2001/: Statistics Canada Symposium 2001 — Achieving Data Quality in a Statistical Agency: a Methodological Perspective

web.mit.edu/tdqm: Total Data Quality Management Web site at MIT