



National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709-4006
Tel: 919.685.9300 FAX: 919.685.9310
www.niss.org

NISS Affiliates Technology Day: Data Quality

NISS Headquarters, Research Triangle Park, NC

February 28, 2002

Summary Report

1 Participants

Participating in the inaugural NISS Affiliates Technology Day were David Banks (Food and Drug Administration), Donald Burdick (MetaMetrics), Lawrence Cox (National Center for Health Statistics), Adrian Dobra (NISS), Alan Forsythe (Amgen), Shanti Gomatam (NISS), Tailen Hsing (Texas A&M University), Alan Karr (NISS), Jon Kettenring (Telcordia Technologies), James Landwehr (Avaya Labs Research), Russell Lenth (University of Iowa), Brent Pulsipher (Pacific Northwest National Laboratory), Robert Rodriguez (SAS Institute), Ashish Sanil (NISS), Yves Thibaudeau (Census Bureau), Lynn Weidman (Bureau of Transportation Statistics) and Tommy Wright (Census Bureau).

2 Presentations

Karr and Sanil presented results from exploratory NISS research on data quality (DQ) and plans for future research on DQ. The exploratory studies were for two NISS affiliates:

US Environmental Protection Agency (EPA), in the Spring of 2001, focusing temporal and trend issues for the Toxic Release Inventory (TRI).

Bureau of Transportation Statistics (BTS), in the Fall of 2001, addressing DQ for the Intermodal Transportation Database (ITDB), with particular emphasis on assessment of DQ.

The presentations were prefaced by a summary of the NISS Affiliates Data Quality Workshop held in November, 2000.¹

Plans for future research included recommendations to EPA and BTS and the NISS proposal *Data Confidentiality, Data Quality and Data Integration for Federal Databases: Foundations to*

¹The workshop report is available at www.niss.org/affiliates/dqworkshop/report/dq-report.pdf.

Software Prototypes submitted in July, 2001 to the NSF Digital Government (DG) program. This proposal (a project summary is attached), which involves multiple affiliates as partners or collaborators, addresses DQ in conjunction with two other urgent problems facing Federal statistical agencies — data confidentiality (DC) and data integration (DI).

PDF versions of the presentations are available at www.niss.org/affiliates/dqtechday200202.

3 Points Raised in Discussions

Discussion took place throughout the presentations, and not all of it can be reproduced here. However, the main points were captured and organized.

3.1 Consensus Points

In general, these confirm findings of the DQ Workshop.

1. DQ is contextual and multi-dimensional.
2. In assessing DQ, talking to users is essential.
3. Data generation processes have profound impact on DQ, and human elements of these processes (e.g., filling out the Form R used to collect data for the TRI) are central.
4. DQ cannot sensibly be addressed in the absence of cost considerations, both to improve DQ and to quantify the impact of low DQ. Costs are difficult because of multiple stakeholders — data assemblers, data disseminators and users, among others.

3.2 General Issues and Questions

1. Differences between active and passive (i.e., legacy) databases.
2. What does DQ *not* include?
3. Other than the scale of the problem, has DQ advanced over the past thirty years?
4. Documentation of many databases is weak, with detrimental impact on usability. Can XML help in this regard?
5. Systems for automatic “cleaning” of data are highly desirable, but are they possible?
6. What does DQ mean for databases that are not predominantly numerical, such as those containing textual or multimedia (in particular, biometric) data?
7. Can user workarounds to deal with DQ problems be captured, and made available to other users?

8. Use of sampling and hypothesis tests as means of validating data.
9. Recently promulgated Federal procedures for information quality were noted; these are available at www.whitehouse.gov/omb/fedreg/final_information_quality_guidelines.html.

3.3 Issues for NISS to Pursue

1. A cost-benefit analysis of a specific DQ improvement.
2. Models for the probability that a datum is correct as a function of cost, producer incentives,
3. Rating systems (e.g., letter grades) for DQ, as opposed to DQ metrics.
4. As in the DG II proposal, mathematization of the effect of DI on DQ. However, separation of DQ from DC was felt to be desirable when possible and appropriate.
5. A review paper on DQ, based on the two exploratory studies.
6. A short course on DQ.

PROJECT SUMMARY
Data Confidentiality, Data Quality and Data Integration
for Federal Databases:
Foundations to Software Prototypes

This is a proposal for a large-scale, cross-disciplinary, high-impact research program to create abstractions, theory, implementable methodology and software prototypes to meet three central, interacting, data-driven challenges facing Federal statistical agencies — DC, DQ and DI.

Federal government-unique problems — especially the necessity, in a electronic world, to balance privacy and confidentiality against user access to high-quality statistical data — define the research. The project will create effective, credible ways to ensure DC in the face of strong, even competing, concern about DQ and the growing need and capability for DI.

The project addresses fundamental research questions in multiple disciplines: *computer science*, to formulate abstractions and design algorithms that accommodate interactions among DC, DQ and DI; the *statistical sciences*, to provide decision-theoretic formulations that account for both the risk and the utility of disseminating information, and of the consequences of DC, DQ and DI for inference; and *software and systems engineering*, to build prototype systems that operate at realistic scales, in order to evaluate and refine new theory and methodology. Complementing these are *domain knowledge*, to link uses of information to requirements for DC and DQ; and *visualization*, to support understanding of abstractions, algorithms, and system operation.

Scalability pervades the research: many techniques for ensuring DC, improving DQ and performing DI are untried, and must be evaluated, in the context of the size and dimension of the databases, the diversity of user needs and the complexity of analyses that we will address. One central challenge is to build systems that correctly implement solutions to technical problems *and* scale to the large, complex databases maintained by Federal statistical agencies.

The project will be carried out by statistical and computer scientists from the National Institute of Statistical Sciences, Carnegie Mellon University, the University of Maryland College Park, the Institute for Social Research at the University of Michigan, Purdue University, Southern Methodist University and the Los Alamos National Laboratory.

As partners in the project, five leading Federal statistical agencies — the Bureau of Labor Statistics, the Bureau of Transportation Statistics, the Census Bureau, the National Agricultural Statistics Service and the National Center for Education Statistics — will ensure that the research is relevant, timely and applicable. The partners will provide essential access to data and participation of personnel in development and evaluation of methods and software systems.