Topic

The devil is in the data

Expect challenging data issues when integrating information

The devil is in the data

Expect challenging data issues when integrating information

Many agencies recognize the benefits of sharing information across programs and reusing existing data resources to provide citizens with integrated services. But they run face-first into data challenges when creating systems to realize these benefits.

Key Points

Data challenges include knowing...

• the devil is in the data

• there is more to data quality than just clean data

• data needs to be 'fit for use'

• meta data is critical

• data content is as important as data quality

Turning data into actionable information

The data challenges agencies face when integrating information and services are daunting. In the world where agencies operate, they must bridge the gap between their business challenges (i.e. program initiatives) and the "relevant" data available to support them. Agencies look to data as the raw material for decision making and planning for the foundation underneath actions taken by the agency.

Turning data into actionable information requires an understanding of what must be done and the data necessary to do it. Lakshmi Mohan, associate professor from the School of Business at the University of Albany, states the key is to focus on determining the difference between what is a "*must do*" versus a "*nice to do*" when it comes to designing an information resource. Organizations must focus first on what must be done and then on finding the relevant data resources to do it. Determining the heritage of the data, assessing its timeliness and quality are all critical and complex parts of the process of turning data into "*actionable information*."

Unfortunately, many policy and program initiatives falter or fail because these challenges are overlooked or are overwhelming. Unexpected levels of effort are often required to:

Focus on "must do and know"

Lakshmi Mohan, Associate Professor
School of Business, University at Albany

- identify relevant data
- determine its usability
- address inaccurate or incomplete data
- deal with the inability to solve certain data problems
- identify and manage confidence in the resulting new resources

All organizations face these challenges. The participants in the *Using Information in Government Program* found the challenges fell into four categories:

"When you look at information you ask the question 'so what'?"

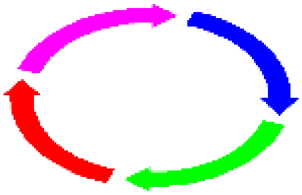
Lakshmi Mohan, Associate Professor
School of Business, University at Albany

- data quality and fitness for use
- data standards within individual systems, as well as across integrated systems
- meta data
- contextual knowledge of the programs where the data is created and used

Fitness for use—the data quality challenge

The raw material of the information age is data. The quality of data comprises its accuracy, completeness, timeliness, relevance, and interpretability in the context of its "fitness for use." In other words, is the quality of the data "good enough" for its intended purpose.

Fitness for Use Cycle



Does the data set **contain** the necessary data elements to answer the business question?

Are the data **relevant** enough given the intended use?

Are the data **accurate** enough given the intended use?

Are the data **complete** enough given the intended use?

Are the data **timely** enough given the intended use?

If an answer is **no**, then what will it take (time, cost, effort) to make it fit for the intended use? Are we willing to pay the price?

If **not**, are there alternative data sources?

If **not**, are we willing to incur the cost to create the data source?

If **not**, should we change the business question to match the data?

With each new data set you must start the cycle over.

Giri Tayi, associate professor from the School of Business at the University at Albany, asserts that data quality management means different things to different people depending on their perspective.

From the analyst′s perspective data quality management requires:

- a sound understanding of the nature of data
- identifying the factors that determine its quality
- defining the costs associated with "good enough"

From an organizational perspective, data quality management means insuring quality commensurate with the various uses of data through:

- specifying policies
- identifying techniques
- establishing quality control procedures

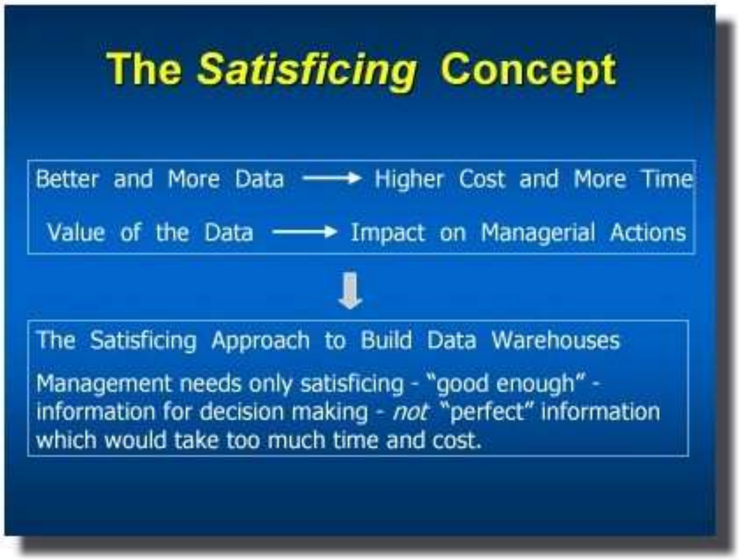
In practice good data quality management demands both perspectives.

Improving the quality of data is costly and time consuming. Organizations must consider these costs in the context of the intended use of the data to determine if the costs are warranted. A number of steps must be taken before a review of costs can take place. Organizations must first:

- develop a full understanding of what data are required to support the intended use
- determine if the available data meet those requirements

If the available data does not meet the requirements, then:

- identify the steps to make the data fit for use by addressing data quality issues or identify the steps involved in acquiring new data
- review the costs of taking those steps
- decide if the costs are warranted



The *Satisficing* Concept

Better and More Data → Higher Cost and More Time

Value of the Data → Impact on Managerial Actions

The Satisficing Approach to Build Data Warehouses

Management needs only satisficing - “good enough” - information for decision making - *not* “perfect” information which would take too much time and cost.

Being clear about what is "good enough" is essential. In order to make reasonable decisions about investments in resolving data quality issues, project managers must define for their organization the difference between "perfect" and "good enough." Since each action or outcome has a cost associated with it, organizations need to decide if the available data are "good enough" for the task at hand. And they need to realize that each notch up the scale toward "perfect" costs time, money, and opportunity.

In the *Using Information in Government Program*, we found these general data quality rules, formulated by Orr (1996), to be useful:

- data that are not used cannot be correct for very long
- data quality in an information system is a function of its use, not its collection
- data quality will not be better than its most stringent use
- data quality problems tend to become worse with the age of the system
- laws of data quality apply equally to data and meta data
- variations among the data sources: attitudes, policies, and practices contribute to uneven data quality

Additional information regarding data quality management issues can be found at:

- *Data Quality Tools for Data Warehousing - A Small Sample Survey*—A CTG research paper discussing several data quality tools.
- ***Data Quality and Systems Theory***—A research paper discussing how data quality, meta data, and systems theory work together.
- *Total Data Quality Management Research Program at M.I.T.*
- ***DATA QUALITY***—An annual, review journal and newsletter
- Additional links

Common ground—the data standards challenge

Information collected by state and local government agencies is a valuable resource. Thousands of files, databases, and data warehouses have been developed. But they aren′t always compatible and in many cases contain duplicate information. These limit our ability to share and integrate information.

The lack of common data standards across these various systems creates a significant barrier to information use. The challenge of creating and implementing unified data standards is compounded when the effort to use information spans organizational boundaries. Creating data standards within this environment requires:

- identifying what data models are used in each organization
- assessing the extent to which they used the same approach, let alone the same elements
- describing specific situations or elements
- determining if there is overlap
- collaborating to develop a "meta-standard" that can be used to guide integration of multiple sources from multiple organizations

Additional information about data standards development and use issues can be found at:

- The **United States Department of Agriculture—Service Center Initiative**—This provides many examples of good data standards and meta data policies.
- NYS Office for Technology’s Geographic Information Systems (GIS) Data Sharing Policy
- NYS Office for Technology’s Data Sharing Policy
- Additional links

Information about information—the meta data challenge

Information about the data—or meta data——often contains:

- a data set's history

- information on how it has changed over time
- what specific rationale guided those change decisions

This information often resides in the heads of those who were involved in the creation of a particular information resource. They know why a data set was created, what rules governed the creation, who the intended users were, and what it shouldn′t be used for. The creators of the data set may not have written this information down or shared it with others because the value at the time was limited to that particular program or situation. But when others try to use a specific data set outside the confines of the original program, the need for good meta data becomes painfully clear. The information required to guide fitness for use decisions, to determine standards used in data collection, and many other questions about the potential value of a data resource is often unavailable. This leads to unused, or unknowingly misused, data resources.

As efforts to integrate data from across multiple programs and governments are increasing, appreciation for the critical role that meta data performs is growing. Meta data can provide knowledge about the fitness for use of a particular data set for a specific decision or assessment. Meta data are not always available or required in the initial implementation of a stand-alone system. But systems that try to integrate multiple data sets without explicit meta data will be, at best, delayed, and at worst, derailed due to the high cost of creating meta data after the fact.

More information about meta data can be found at:

- Meta Data Standards and Registries: An Overview
- Dr. Tom's Meta-Data Primer—Provides a basic understanding of meta data.
- The NYS GIs Clearinghouse and Geographic Information System Data Cooperative
- An Introduction to Metadata (from the Getty Research Institute)
- Additional links

Understanding the program environment—the contextual knowledge challenge

Government data are usually a reflection of something else—perhaps a program, a service, a person, or a process. Anyone who uses the information needs to know about its context in order to use it well. But this knowledge is not always available in the form of explicit meta data, because meta data standards generally do not require this type of information. It usually resides in the working knowledge gained through years of managerial experience of managing those programs and services, not with the technologists who develop systems.

Program managers are often involved in the initial discussions regarding the functional design of a proposed system, but are not involved in any subsequent system processes until the system is ready for use. Without their involvement, data inclusion and exclusion decisions can be incorrect. Like meta data, contextual knowledge is important to avoid misuse of the data. All relevant program managers need to be involved in the process of deciding fitness for use. Their knowledge of what the data actually represents is crucial when developing systems that utilize existing data or data obtained from outside sources.

Additional information that illustrates these challenges and offers techniques for addressing them:

- Research and Practical Experiences in the Use of Multiple Data Sources for Enterprise Level Planning and Decision Making: A Literature Review
- "Dealing with Data Seminar Summary Report"—A CTG report that addresses a variety of data issues specifically fitness for use.
- "Putting Information Together: Building Integrated Data Repositories Seminar Summary Report"—A CTG report that discusses the issues organizations face when integrating data from multiple sources.
- Additional links

Practical Examples

The following examples from recent data integration projects show how agencies are dealing with data issues:

Integrating disparate data sources

The challenges of data quality, data standards, meta data, and contextual knowledge exists in any information use initiative. These challenges are compounded with each additional data source. Agencies building new integrated data repositories by drawing data from different government programs must recognize and plan for these challenges. The Homeless Information Management System (HIMS) at the NYS Bureau of Shelter Services (BSS).

Managing differences in the data

Facing a difficult task of working with a variety of organizations to build an integrated information repository, the Kids Well-being Indicators Clearinghouse (KWIC) team initially designed the Clearinghouse to support the broad range of queries requested by users. However, the data available from the various sources often were unable to support the cross comparison of data sets that users wanted.

For example, population figures from different sources were used in different calculations. Because of this, the ability to compare across variables was more limited than expected. The limitations of the data could not be overcome by mere technology; it would have required costly changes in each of the underlying data sources. Without program staff who were knowledgeable about the population sources, an error could have been made by allowing unsuspecting users to make invalid comparisons. It was important that the KWIC development team have strong background knowledge about the data, understand its complexities, and manage the data differences.

Determining "fitness for use"

Data quality problems became a major barrier to the usability of the available data in the Homeless Information Management System. The team needed to determine the extent of the data quality problems, and their ability to mitigate them, in order to determine if the data was fit for use. The Homeless Information Management System depends on the ability to integrate a large number of highly disparate sources of data collected under a variety of different circumstances. Each provider, for example, has its own approach to managing the data collection process, therefore each has a different set of minimal standards of acceptability. Identifying and determining the specific data quality issues that needed to be addressed in order for the data to be included was a major part of the development process.

The issues of data relevance and fitness for use took on a different meaning in the project on Assessing IT Investments at the NYS Department of Transportation. This project involved the lead agency broadening its view of the kind of information needed to make these decisions. It not only had to identify the new data elements, but also had to define and specify rules for others to follow when correcting the elements.

Building meta data

Meta data, whether in someone′s head or written down, are essential for system design teams. In the Kids Well-being Indicators Clearinghouse, the meta data needed to contain the following information: why and how the data was collected, and in the case of aggregate data, how the aggregates were calculated. Using one data set as the blueprint, other data sets needed to be augmented to contain similar information. The goal was to have all data be of value for the intended users of the Clearinghouse.

In the Homeless Information Management System project data were integrated and aggregated based on business rules that were specific to the new repository. The meta data from the existing systems were used as the foundation to determine if a data set could be included. New meta data were created to: document the data source, show how the data were aggregated and changed, and define the meaning of the resulting new data.

Communicating the context of the data

The Bureau of Shelter Services team found that contextual knowledge of the data was critical to the success of their prototype integrated data repository. Knowledge about the programs and the clients represented in the data provided a perspective unavailable in the data dictionary or in the minimal meta data.

Both technical and program staff had to actively participate in the decisions to include or exclude specific data sets. Program staff provided the legal, historical, or operational rationale for the use of particular data elements and codes. Without this kind of contextual knowledge, important distinctions in the data would have been lost.

Data Links

Data Quality Tools for Data Warehousing - A Small Sample Survey
This paper, generated by the Center for Technology in Government, discusses how data quality is one of the biggest issues people face when integrating data. Data quality tools are used in data warehousing to ready the data and ensure its cleanliness. This research focuses on how the data quality tools address problems in data.

Data Quality and Systems Theory
This paper, written by Ken Orr from the Ken Orr Institute, discusses how data quality, meta data, and systems theory influence information systems and the results they generate.

Total Data Quality Management Research Program at M.I.T.
The overall objective of this program is to establish a solid theoretical foundation for Data Quality Management to devise practical methods for business and industry to improve data quality. It includes tools and other capabilities necessary for data quality management in the technical, economic, and organizational phases of business operations. Also listed on this site are links to conferences and papers about data quality.

DATA QUALITY "is an annual peer-reviewed journal founded in 1994 by a group of statisticians, information scientists, and quality practitioners. DATA QUALITY includes original research papers and book reviews about data and information quality. They range in scope from general interest to theoretical."

Tools for Traveling Data
This article, written by Joseph Williams, discusses the idea that "one of the most important but overlooked steps in building a data warehouse is loading data in the

warehouse database. This article discusses three broad categories of tools for helping load data into a data warehouse-data quality, extraction and transformation, and cleansing."

Introduction to Metadata: Setting the Stage
Metadata, literally "data about data," is an increasingly ubiquitous term that is understood in different ways by the diverse professional communities that design, create, describe, preserve, and use information systems and resources. This article provides a description of the different uses and categories of metadata and their various roles and functions.

Technology Policy 97-6 Geographic Information Systems (GIS) Data Sharing
This document, set forth by the NYS Office for Technology, discusses how electronic geographic data can be shared among federal, state, and local agencies. It reviews provisions that have been made to ensure that GIs data is shared in an appropriate and consistent manner. Custodianship, standards, maintenance, pricing, requests, and submissions are outlined.

Technology Policy 96-19 Data Sharing Among Agencies
This document, set forth by the NYS Office for Technology, describes steps that NYS agencies should follow when considering sharing data electronically. In addition, this policy helps ensure a standard process of data sharing among NYS agencies. Each step describes a key component in data sharing and includes questions that can help an agency prepare and complete that step. Also included is a sample case study that describes one agency's initiative through each step.

Technology Policy 97-3 Statewide Data Dictionary
This policy, set forth by the NYS Office for Technology, announces the establishment of a "Statewide Data Dictionary" that features a core section of data elements that cross all State agencies (i.e., name, address, etc.). It has a section dedicated to each functional area of government activity (i.e., criminal justice, health, etc.) that covers the data elements unique to that area.

Meta Data Standards and Registries: An Overview
This paper (in PDF format), generated by the US Environmental Protection Agency and the US Bureau of Labor Statistics, discusses the work that is being done to reach consensus on standardizing meta data and registries for organizing that meta data. It provides information about meta data and also goes in-depth on the impact a meta data registry can have on a statistical agency.

Dr. Tom's Meta-Data Primer

This paper, written by Dr. Thomas Wason, provides a basic understanding of meta data, how it is structured, what it means, and how it is represented. It also describes when there is sufficient meta data, and how meta data is scaleable and interoperable with other systems.

Creating a Statewide Spatial Data Repository and Geographic Information System Data Cooperative

The NYS GIS Clearninghouse shows data sharing was integral to the successful creation of the NYS GIS Clearinghouse. This project focused on tools and policies for data sharing. A prototype meta data repository accessible over the Internet was produced. It included an inventory of spatial data resources around the state, and a set of policy and management recommendations for a permanent data cooperative.

An Introduction to Metadata: Pathways to Digital Information

This site is a collection of metadata articles, a suite of standards, and useful links for Metadata for the World Wide Web. This metadata resource strives to "help those with a stake in the debate (everyone from librarians to museum professionals to anyone who intends to make information available via the Internet) to avoid mistakes and wasted effort, and to make informed decisions about the information they seek to record and disseminate." The Getty Research Institute, which sponsored the development of this site, "promotes innovative scholarship in the arts and humanities, to bridge traditional academic boundaries, and to provide a unique environment for research, critical inquiry, and debate."

Research and Practical Experiences in the Use of Multiple Data Sources for Enterprise Level Planning and Decision Making: A Literature Review

This paper written by the Center for Technology in Government, discusses how many agencies are faced with the challenge of integrating data sources to provide a broader look at programs and service. To do this, information sharing is essential. It allows agencies to improve planning and increase productivity. The use of multiple data sources for enterprise level planning and decision making is also increasing. This paper identifies current research. It also outlines practical experiences in the use of multiple data sources to support performance measurement, strategic planning, and interorganizational business processes.

Dealing with Data Seminar Summary Report

This report (in PDF format) from CTG's seminar "Dealing with Data," addresses many data issues through summaries of presentations on data quality management, data tools and techniques, long term maintenance and preservation, and other data issues.

Putting Information Together: Building Integrated Data Repositories Seminar Summary Report

Using CTG's Homeless Information Management System prototype project as an example, this seminar summary report discusses the management, policy, and technology issues organizations face when integrating data from multiple sources, to create a new information resource for cross program, cross organizational decision making and planning.

DM Review

DM Review is a monthly issues and solutions publication that focuses on data warehousing and business intelligence.

DB2 Magazine

DB2 Magazine is an electronic publication dedicated to providing strategies and solutions for database programmers, administrators, and users.

Data Warehousing: Introduction to data warehousing

A web-based resource produced by the Operational Research Society to help members of the OR community gain a good understanding of data warehousing. Efficient data warehousing is seen as crucial in enabling companies and organizations to utilize existing information by providing a a central data repository of stable, accurate, consistent and clearly understood data.

Turning Data into Understanding: A Field Guide to Knowledge Support Technology

The guide (in PDF format) developed by the NYS Forum for Information Resource Management, is designed to help government decision makers determine whether technology tools can and should be used to better understand the problems facing their organizations.

Keyword Search Suggestions

- "Data Integration"
- "Data Mining"
- "Data Standards"
- "Data Warehousing"
- "Data Cleansing"
- "Data Extraction"
- "Data Migrating"