

## **The devil is in the data: Expect challenges when integrating data for e-government**

---

### **Introduction**

Many agencies recognize that e-government can provide the benefits of sharing information across programs and providing citizens with integrated services. But those same agencies run face-first into data challenges when creating e-government systems to realize these benefits.

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 an e-government application 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:

- 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. They often fall into four categories:

- 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

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 our research and practice, we have 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

---

## **Common ground—the data standards challenge**

Information collected by state and local government agencies can be a valuable resource on which to build e-government programs. 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

---

## **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.

---

## **Understanding the program environment—the contextual knowledge challenge**

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 experience managing those programs and services, not with the technologists who develop e-government systems.

Center for Technology in Government/University at Albany, SUNY  
Mark LaVigne, (518)442-4598, mlavigne@ctg.albany.edu

Program managers are often involved in the initial discussions regarding the functional design of a proposed e-government 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.