Data Quality: Letting Data Speak for Itself within the Enterprise Data Strategy

Collaboration & Transformation (C&T) Shared Interest Group (SIG)

Financial Management Committee

DATA Act – Transparency in Federal Financials Project

Date Released: September 2015

SYNOPSIS

Data quality and information management across the enterprise is challenging. Today's information systems consist of multiple, interdependent platforms that are interfaced across the enterprise. Each of these systems and the business processes they support often use and define the same piece of data differently. How the information is used across the enterprise becomes a critical part of the equation when managing data quality.

Letting the data speak for itself is the process of analyzing how the data and information is used across the enterprise, providing details on how the data is defined, what systems and processes are responsible for authoring and maintaining the data, and how the information is used to support the agency, and what needs to be done to support data quality and governance. This information also plays a critical role in the agency's capacity to meet the requirements of the DATA Act.

American Council for Technology-Industry Advisory Council (ACT-IAC)
3040 Williams Drive, Suite 500, Fairfax, VA 22031
www.actiac.org ● (p) (703) 208.4800 (f) ● (703) 208.4805

Advancing Government through Collaboration, Education and Action

American Council for Technology-Industry Advisory Council (ACT-IAC)

The American Council for Technology (ACT) is a non-profit educational organization established in 1979 to improve government through the efficient and innovative application of information technology. In 1989 ACT established the Industry Advisory Council (IAC) to bring industry and government executives together to collaborate on IT issues of interest to the government.

ACT-IAC is a unique, public-private partnership dedicated to helping government use technology to serve the public. The purposes of the organization are to communicate, educate, inform, and collaborate. ACT-IAC responds to government requests using a model that includes government and industry working together, elbow-to-elbow. ACT-IAC also works to promote the profession of public IT management. ACT-IAC offers a wide range of programs to accomplish these purposes.

ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of IT. For membership and other information, visit the ACT-IAC website at www.actiac.org.

Collaboration & Transformation SIG Financial Management Committee

DATA Act – Transparency in Federal Financials Project

The C&T SIG sought input from the Department of the Treasury and the Office of Management & Budget (OMB) to follow the progress of the Digital Accountability and Transparency Act (DATA) Act from the pilot phase through practical/production implementation, providing useful information for industry and government managers to consider as they assess their readiness and develop their strategies to meet the new requirements.

Disclaimer

This document has been prepared to provide information regarding a specific issue. This document does not – nor is it intended to – take a position on any specific course of action or proposal. This document does not – nor is it intended to – endorse or recommend any specific technology, product or vendor. The views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development. Every effort has been made to present accurate and reliable information in this report. However, ACT-IAC assumes no responsibility for consequences resulting from the use of the information herein.

Copyright

Further Information

For further information, contact the American Council for Technology-Industry Advisory Council at (703) 208-4800 or www.actiac.org.

Executive Summary

Most organizations, both in government and private industry, are concerned about the quality of their data. They know that data quality issues often present themselves in different ways. They may see the signs of questionable quality during month-end close and financial audits. They may see the user's confidence in the data and systems degrade. The challenge is not in knowing if there are data quality concerns, the challenge is knowing exactly what the data issues are, how to fix them, and where to start.

Understanding and managing data quality is a daunting task in any organization. The business process complexities, antiquated systems, complex requirements and mandates, and the sheer volume of data within the federal government can make this seemingly impossible. To make an already difficult situation worse, different departments within the agency often use the same piece of data differently, and the definitions of those data elements may even be in conflict with each other.

To fully understand the quality of data, you must first understand how the data that is currently in the systems is being used to support current processes, and what the business is doing outside of the systems when they encounter data quality issues. Letting the data speak for itself does just that.

Letting the data speak for itself is done by conducting fact-based analysis on all of the agency data. It includes analyzing master and transactional data, how the master data supports the specific transactional data, and the data relationships across systems and business processes. Agencies are currently using the data to conduct their operations. The data is defined by the business to support the business processes. Aligning this analysis and the data with the business processes it supports identifies the gaps in data quality, and the business process workarounds being used to compensate for poor data quality.

Every agency must be able to provide accurate, high-quality data for its implementation of the DATA Act to be successful. The DATA Act is not just about standards, it is also about transparency. In addition to high quality data, the agency leadership must understand how the data is being used by the business. Without that understanding they will find it difficult to respond to questions about the quality and usefulness of the data.

Letting the data speak for itself identifies potential gaps in data quality and related business processes and provides understanding of how to improve data quality for the benefit of the agency, and for the successful implementation of the DATA Act.

Introduction

As the acronym for the Digital Accountability and Transparency (DATA) Act implies, data is at the heart of everything we do in business and in government. In fact, data is the "DNA" of our organizations, for it is embedded in every business and government process and informs the organization's mission and how it functions. Analyzing an organization's data across the enterprise can provide critical insights into business and operating requirements and how the data is defined and consumed – as well as what must be done to satisfy DATA Act

requirements. As such, ensuring that you have quality data is crucial not only to successful DATA Act implementation but perhaps also to the success of your organization.

What Does "Data Quality" Mean?

Many definitions exist for data quality, one such example being "The state of completeness, validity, consistency, timeliness, and accuracy that makes data appropriate for a specific use."[1] As most definitions include these or similar terms, data quality obviously must address completeness – the data value must reflect all of the information it was designed to capture or convey, validity – the data comes from a reliable source or process, consistency – the same source or process should produce the same data and the same data values for a given event or object should be reflected across the enterprise, timeliness – data should be sufficiently current for use, and accuracy – the correct value should be recorded at the point of inception and this value should be retained across the enterprise. The measurement of data quality should be easily available to consumers of information so that they may determine fitness for a specific use in support of fact-based decisions. The accuracy of data is more important for high-risk decisions than for low-risk decisions, and factors into the decision process.

Another often overlooked aspect of understanding data quality is latency. The context and source of each piece of data and the timing of when the data is provided is extremely important. If the latency of the data and its source is not understood it can cause perceived inconsistencies when the data itself is accurate – the data is not wrong; it is just of a different timeline. For example, the same data element published on USASpending.gov and another website can be different if the data was sourced at different times, or if the data was sourced from different interfaced systems that are not updated in real time.

The "Ins" and "Un" of Quality Data

The data quality concepts all seem easy enough, and we intend to design business processes and systems that generate and record data in a complete, valid, consistent, timely, and accurate manner – but often this is not the result. Instead, our enterprises generate and possess data that is often incomplete, invalid, inconsistent, inaccurate, and untimely. Why does this occur, especially in government, if our intention is the opposite?

Antiquated Systems

Data is often interfaced across multiple systems and departments and the data, as well as the systems that support it are in a constant state of change. It is not practicable to update every related system when one component requires an upgrade. As individual systems are upgraded or replaced the government must continue to use and maintain the legacy systems that still meet the needs of the agency and its business processes. Although these systems continue to meet the business needs, they often run in mainframe or antiquated settings using outdated programming languages (e.g., COBOL), their continued operation and ability

[1] Government of British Columbia (http://www.oag-bvg.gc.ca/internet/English/att_20030402xe01_e_12685.html), as cited in "Data quality," Wikipedia (http://en.wikipedia.org/wiki/Data_quality).

to interface with other and newer systems is troublesome at best and problematic at worst. Because their logic is often "hard coded," with the concurrent limited ability to accommodate updates and change, embedded computing processes often generate inaccurate data that must be fixed via manual workarounds. Further, their limited ability to interface with other systems requires the manual transfer of data, which introduces further opportunities for error. Until these legacy systems can be fully retired and subsumed by modern architecture, intensive manual efforts will continue to be required to manage the resulting data, with quality continuing to be challenged as a result.

Federal Unfunded Mandates

Congress often passes new laws and modifies existing standards to help improve the availability, standardization, transparency, and accountability of data across the government. Presidential Administrations also release directives, policies, and executive orders that, although beneficial for positive change, are often difficult to act on, given the agencies numerous other mission critical priorities and budgetary constraints. Many of these changes mandate new or expanded activities that often come without additional funding. As statutes and regulations change, implementing business processes also must change – and if a system helps operate the business process, it too must change. However, as discussed above, legacy applications can be very difficult to reprogram, and even modern enterprise resource planning-based systems can have trouble keeping up with ever-changing requirements, not to mention having to implement new rules retroactively depending on how fast implementing guidance is issued, and these changes often must be funded out of current agency resources. All of this may mean more manual workarounds, rushed implementations, and the introduction of more opportunities for poor data quality.

Continually Changing Standards

As laws and regulations change, so do our IT standards. While continuous improvement in automated system governance standards is laudable, it can be challenging for government to achieve compliance. Identifying additional resources to comply with the newest standard is becoming increasingly difficult in an era of sequestration and austere budgets. In addition, standards are often ambiguous, leaving federal organizations, many of which lack the expertise in IT standards and governance, unsure of the appropriate path forward.

Departmental Silos

Federal departments and agencies are experts at their mission and what they do to support our nation and citizens. Many are also experts at operating within their own silo, but often have limited visibility to the data definition and use outside their business processes, and thus miss the opportunity for collaboration outside of their office on many issues, data sharing included. This practice hinders efforts to reconcile data across disparate systems and offices, resulting in inconsistent data that often goes undiscovered and unaddressed.

The Executive Order and accompanying Open Data Policy of May 2013 was designed to assist agencies towards opening their "silos of excellence," and some agency management's

performance is based on their sharing of data. These mandates ensure a higher data quality and consistency than we have seen before.

Reactive Data Validation/Quality Checks

Quality checks commonly occur at the end of the business process lifecycle, executed by consuming or receiving applications. Errors are identified and the data is sent back to the producing application owner for assessment and correction. This reactive process results in poor quality data being passed back and forth between producing and consuming applications – until all of the errors are identified and subsequently corrected, often by the producing application owner. The receiving application's executed validation takes time and increases the cost of data quality as the effort to identify and correct the issues commonly occurs when consumption and use is imminent, and requires recycling to correct errors identified by the receiving application's validation checks. This can increase the risk to data quality as this reactive process often requires the data to be remediated out of the context of the business process that created the data.

What Does This Mean?

The cumulative effect of incomplete, invalid, inconsistent, inaccurate, and untimely data is significant. First, poor data quality hinders the ability of leaders and managers to adequately assess risk, set priorities, and manage their operations effectively. Second, the need for large-scale manual workarounds, and the negative mission impacts that poor data can cause, drive the use of more organizational resources – time and/or funding – than would be required if the data quality was high. The organization would reap more benefit from applying those resources to other needs, such as data analytics, that would provide insight not only into fulfilling the agency's mission, but also in how to deliver the mission more effectively. And third, inaccurate operating and performance data can lead management into poor decision-making. For example, a program that is tracked as performing higher than it actually is may cause managers to miss key warning signs that the program will miss targets; the opposite might encourage managers to divert resources needlessly to the program that could have been better spent elsewhere. As described in greater detail below, poor data quality is a large drag on efficient and effective mission accomplishment, making a sound business case for efforts that will improve data quality.

Background

Understanding Data Quality and What to Do About It

In every organization, data is needed to perform key business functions such as program management, budgeting, resource allocation, strategic planning, and to provide insight into improving performance; hence, it is essential to understand the quality of the data. Fundamental steps in data management are assessing, improving, and monitoring data quality. Every federal agency should be able to answer the questions, "What is the quality of the agency's data?" and "How is data presented so that users can understand its quality and determine whether it is fit for its intended use?"

Learning the Quality of Your Data

How much an agency knows about the quality of its data largely depends on how much time and effort it invests in data quality assessment, improvement, and monitoring, but they also learn about the quality of their because a data quality issue hinders day-to-day operations. For example, when financials do not reconcile during fiscal year close, little time is invested to discover the source of the problem, such as a data entry error, incorrect formula, or improper relationship between related relationships. Not only is time required to discover the source of the data quality issue, but often the fix requires time-consuming manual workarounds that are susceptible to introducing additional errors into the data.

In less ideal cases, agencies learn of errors in their data from external users. One such example is the review of federal data for audits or GAO studies. GAO Comptroller General Gene Dodaro said in reference to the use of agency data for performing studies, "If the data quality is not good, you're limited in your abilities to use it. In many cases, it takes us a lot of time and effort to assure ourselves of the quality of the data when, in my belief, it should be assured by the executive departments and agencies as a routine manner as a management responsibility."[2]

GAO has published numerous studies that directly address a federal program's data, citing cases in which poor data quality has impeded an agency's ability to function effectively.[3] In another example not related to GAO, one agency discovered that the data it released was inaccurate when the data was used to identify the best high schools. The principal of a high school ranked 13th in the nation in 2012 questioned the honorable ranking when U.S. News and World Report showed the school as having a 4:1 student to teacher ratio instead of the actual 24:1 ratio.[4] The news magazine reported that it had pulled the data from statistical data the agency makes available to scholars, media, and the public.[5]

The adoption of a data quality management system enables agencies to detect and address data anomalies prior to releasing data, keeping user confidence in the data intact. Following well-established data quality improvement processes, agencies can confidently comply with the requirements of the DATA Act and the Data Quality Act. Such processes provide an objective and methodical way to identify data anomalies, determine whether they are valid errors, address those errors (either through correcting the error at the source, or attaching an annotation to the data regarding the error), and monitor the data quality and the data quality trends.

---

[2] Kopp, Emily. (2014, November 13). "GAO: Not too Early to Track DATA Act Progress." *Federal News Radio.* Retrieved from http://www.federalnewsradio.com/440/3741840/GAO-Not-too-early-to-track-DATA-Act-progress

[3] For example, a search for "data quality" on the GAO website produces approximately 1,500 reports and testimonies containing this term.

[4] Takahashi, Paul. (2012, May 9). "U.S. News 'looking into' reports of erroneous data in Best High Schools rankings." *Las Vegas Sun.* Retrieved from http://lasvegassun.com/news/2012/may/09/us-news-looking-reports-erroneous-data-best-high-s/

[5] Ibid.

A few federal agencies are already realizing the benefits of deliberately and proactively addressing data quality. Representatives from the Consumer Financial Protection Bureau and the Federal Reserve Board of Governors spoke at the Data Transparency Coalition March 2015 Financial Regulation Summit, and noted that data quality standards set by the two agencies have made it possible for them to use each other's data with confidence. Similarly, a GAO report published in July 2012, titled "Government Transparency: Efforts to Improve Information on Federal Spending," documented how lessons learned and implementation considerations, such as data quality, from the Recovery Accountability and Transparency Board (RATB) could be leveraged to provide broader government transparency.[6]

The high quality of stimulus data did not happen overnight. First, recipients and sub-recipients reported their data in a well-documented web form front-end in Federalreporting.gov. In order to get better end results, the RATB included clear instructions and definitions of the data to be entered and developed strong data field validations in the web form. For example, open text fields changed to drop down selections where it made sense and standardized numeric character types were used for fields such as phone and zip code.

In a recent interview, Department of Housing Development (HUD) Deputy CFO Joe Hungate echoed that having mechanisms up front, similar to Federalreporting.gov, in the process leads to stronger data quality. He noted, "The root cause of most bad data is at the entry point." He also mentioned that good guidance on the forms themselves is equally important to ensure that users properly interpret what they should include. Standardized definitions combined with data entry points that can be better managed will contribute to stronger data. Mr. Hungate's closing statement on this point was "Improving the quality at the atomic level improves the aggregate."[7]

Acting on Data Quality

Government agencies are not alone in their reticence towards data quality. A Gartner report[8] on the business value measurement of data quality noted that a majority of organizations do not proactively manage data quality, but deal with the fallout of poor quality data as issues arise. This approach is not sustainable, particularly as organizations increasingly automate their business processes, such as payments and grants management. Data quality defects impact business processes and become a limiting factor to optimizing process quality.

Not knowing the magnitude of their data quality issues or the resources required to tackle them may cause some organizations to defer seeking the benefits of data quality management. Adopting a data quality improvement system means methodically uncovering data quality issues that, once revealed, must be addressed. It means being cognizant of just

---

[6] *Government* Transparency*: Efforts to Improve Information on Federal Spending*, GAO-12-913T: published: July 18, 2012. Retrieved from http://www.gao.gov/products/GAO-12-913T.
[7] Joe Hungate, May 8, 2015 ACT-IAC project team interview
[8] Friedman, Ted and Smith, Michael. (2011, October 10). *Measuring the Business Value of Data Quality.* Retrieved from https://www.data.com/export/sites/data/common/assets/pdf/DS_Gartner.pdf

how good or bad an organization's data is and having to determine what level of quality is good enough. It means dedicating sufficient resources – people, processes, and technology – to efficiently and satisfactorily measure and improve data quality. This can all seem overwhelming, but it does not have to be, particularly given that the real burden and risk is in not knowing the errors that lurk in the data and their potential negative impacts of those errors. That risk is being uncovered by data analytics and the program management teams are starting to understand the value of data quality towards improved program performance. The Gartner report cites poor data quality as the chief reason that 40% of all business initiatives do not achieve their expected positive outcomes.[9]

Several processes and tools exist to support organizations in capturing the benefits of data management. In addition to building data quality processes into data collection, as noted in the RATB example above, agencies should check their existing technologies for data quality tools, since some data warehousing platforms and data staging, integration, and analytics software packages come with data profiling features that are the foundation of data quality assessment. Open source data quality tools with free-use licenses, usually with limited functionality, are also available.

Data quality impacts, and is connected to, all organizations within a federal agency, so it is important to engage agency-wide data governance in adopting data quality processes and standards. The governing body can collectively counsel in determining the metrics to use in expressing the quality of the data, and in setting data quality objectives. It may not be realistic to expect error-free data, but a more feasible and economical goal is optimized data quality – not investing more in data quality efforts than the savings realized from resolving data quality issues. The Gartner report also offers a methodology to quantify the business value of data quality improvement, using business metrics tied to the main functions of an organization and their financial impact.[10]

Ultimately, the user of data, whether internal or external to an organization, must determine whether data is fit for use. Given that the same data may be used by different users for different purposes, in some cases an acceptable level of quality for one user may not be acceptable for another. As they develop and implement their data quality improvement systems and measure data quality, agencies should keep in mind that they do not have to determine whether the data are fit for every use. Rather, consider how to annotate questionable data in a succinct, accessible, and clear way that makes users aware of anomalies and provides sufficient information to determine whether the data is suitable for the user's particular application. Understanding the quality of the data is critical to gaining trusted insight – an often unintended organizational asset.

Let the Data Speak for Itself

Data, the DNA of business, is created by the business, to support the business and its processes. An in-depth analysis of master and transactional data gives you critical insight into

[9] Ibid.
[10] Ibid.

how the data was created, what systems or organizational programs authored the data, and where it is consumed. It tells you how the business is using the data, where it resides, and how the data relates to other data and business processes across the enterprise. Letting the data speak for itself gives organizations the fact-based analytics required to truly assess and manage the quality of their data. It identifies what data is relevant and provides the understanding of why the data is gathered, how it is maintained and manipulated, and the business drivers behind the data.

Letting the data speak for itself gives a federal agency the facts required to identify data requirements and quality issues across the enterprise, remediate those issues, and identify gaps in business process and training. It provides the required analytics to develop a business reference model that accounts for how the data is actually being used across the enterprise. This level of understanding is essential to truly manage the quality of a federal agency's data across the enterprise, to map and transform the data to comply with the DATA Act, and to fully leverage the power of agency data.

The business and IT organizations within a federal agency know the data, so why is letting the data speak for itself important? Data is often used differently in different departments, even when using the same schema or application. For example, one department may be using the "contact name" field as the name of the billing contact for the vendor, while another may be using it as the contact for the recipient of the contract or award. The inherent lack of visibility of the business processes and how the supporting data is stored and used across departmental silos often results in duplicate and inconsistent data within the systems. Any disconnect between who is responsible and who is accountable for authoring and maintaining data results in inconsistent and inaccurate reporting. An in-depth analysis of how the data is used in each business area and across the enterprise is necessary to identify how the business is actually using and storing the data, in addition to gaps and inconsistencies in business processes, definitions, and data quality.

New legislation and requirements drive changes in technology. When a new system is implemented, the data in applications that will not be replaced often requires changes to support integration and enterprise-wide reporting. For example, when an agency migrates to a new financial management system, a complex set of interfaces is often built between the contract writing system and the new finance system to ensure that new awards are tracked and reported accurately.

Changes in business requirements drive changes in data collection, storage, usage, and maintenance. These changes are frequently made on the fly, and more often than not they are made in a silo and result in significant data quality issues. Taking the time and expense to coordinate across the organization, especially when a data governance organization is not in place, is often prohibitive. For example, a user needs to report on the number of contracts that are in different stages (request for information, request for quote, request for proposal, etc.). The user therefore makes a decision to attach the contract number with the stage (RFI, RFP, award, etc.). Although this could provide a solution, there are risks. The average user does not know how the data is being consumed across the enterprise. By making a change

to how the contract number is being stored in the contract writing system, the user has broken the referential integrity between the contract system and the finance system.

No one person has a complete understanding of how data is authored, validated, modified, stored, consumed, and eventually archived across an enterprise, yet this understanding is imperative to implementing any data standard, including the DATA Act. The fact that the DATA Act does not mandate how the data is stored by each agency, instead mandating only the definition and format, makes this understanding even more imperative. Achieving that understanding is commonly accomplished through establishing a data governance organization with business and IT representatives. Much of the data that is required by the data standard does not exist in the required format, and often exists in multiple formats across the enterprise. Many of the difficulties with implementing the DATA Act will not be the standard itself; it will be understanding where to find the required data, how to consistently transform that data, and how to reach and maintain an acceptable level of data quality.

What You Need to Know About Your Data

Business data is typically divided into four categories – transactional, master, reference, and metadata.

Transactional data is dynamic data that represents an actual business transaction. Purchase orders and financial transactions are examples of transactional data.

Master data is enterprise-wide data that supports the transactional data. It is the vendor record that represents the vendor on a purchase order, or the fund or cost center record referenced on a financial transaction. Master data is usually, but not always, limited to non-transactional data. For example, the award record for a blanket purchase agreement (BPA) could be considered master data that supports the items on that BPA.

Reference data is very similar to master data and often used interchangeably. It is typically static data that is referenced by and supports the transactional data. Payment terms on the vendor record or purchase order are considered reference data.

Metadata is data about data, and is a key consideration when analyzing data quality. Metadata encompasses multiple dimensions, many of which are outside the scope of this paper. Some of the key metadata elements from a data quality perspective include the business definition of the data, valid formats, allowed values, authorship, ownership, etc. The type of data and its business use drives the importance of each metadata attribute. For example, for free-form text fields, the only required metadata may be the format of the field and its business definition. However, the attributes required to define and manage specific financial or operational data would be much more detailed.

One of the struggles common to many organizations is not identifying the technical definition of the data and how the data is expected to be used across the organization, it's determining how each piece of data is actually being used in the disparate systems. There is often a large divide between vision and reality, especially with systems that have been assembled piecemeal over decades. This is why it is imperative that a formal data assessment be undertaken as part of understanding the data and where it is located. Then the data

governance organization can map the gap between the reality and the vision for the data. Closing this gap sometimes entails a modification of process. Data in itself is of no value; it is not until that data is used in a process that value is realized. Therefore, process and data governance are symbiotic.

Recommendations

How to Let the Data Speak for Itself – Getting Started

The key to letting the data speak for itself is performing an in-depth analysis of all of the transactional and master data, from all systems in scope from both a business and a technical perspective. There are a few key questions to ask before you begin:

- Are there specific areas where the business is experiencing difficulties? The data exists to support the business. Business process interruptions or delays are often due to poor data quality.
- Are there specific business areas or systems where user acceptance is lacking? User acceptance is often an indicator of data issues.
- Are there specific areas where standard reports need to be "fixed" before they can be used?
- What systems should be analyzed? The authoritative sources for all relevant data should be considered as part of the analysis.

Another key aspect to consider is the context and business use of the data being analyzed and ultimately sent to USASpending.gov. Is the data at the summary or detail level? Is this consistent with the data standard for that element?

This becomes increasingly important to ensure that the consumers of the data understand its context and accurate assumptions and decisions are made when they use the data. For example, assumptions are required when visualizations are made. The data consumer may ask what the spending looks like for their area. An assumption is made when the drill level is decided – showing a map at the state level may give the impression that money is spent evenly across the state.

Decisions and assumptions have to be made about the channels for communicating the data without compromising the data context. This is not possible without fully understanding both the context and use of the source data as well as the intended context of the data standard.

Data Quality Assessment

Data is currently being used by the business to support business processes. Analyzing the transactional and master data provides much needed insight into how the business is using the data and information on the business requirements for the data, including the format. Since quality data is defined partially as data that is fit for use, it is imperative to understand the uses of the data to effectively assess its quality. A data assessment consists of two facets: data profiling, which provides information on data with no context, and data auditing,

which measures compliance to known business rules. Data profiling would, for example, show that an organization has a dozen different ways of representing gender, which is typically one of two values. Data profiling results therefore can identify potential data anomalies which require further analysis. By contrast, a data audit would identify data defects – data that violates a defined business rule (for example, charges to an expired account code).

Letting the data speak for itself by comparing the system and technical metadata requirements to how the data is actually being used across the enterprise provides the fact-based analysis to identify gaps in data quality, training, and business processes, and provides the baseline analysis required for developing a business reference data model. A business reference model is a graphical representation of the relationships between the data. It is similar to what many would call a logical data model but with even fewer technical requirements. The intent of the business reference model is a simple visual model to easily see relationships – what entities within the organization use the data, how they use the data, and to what end. It is a great tool for the data governance organization to easily perform impact analyses related to a data element.

Data Quality and Business Relevance

There are differences between error-free data and business-ready data. Having error-free data means that the data is not expected to cause an error within your systems. It also means the technical system and database requirements are met. Error-free data does not mean that there will not be data related business issues or interruptions due to how the data is used in context. For example, a general ledger account can be configured to require or not require a cost assignment when posting to that account. Either case may be valid depending on the business requirement. However, this does not mean that either case will support the business process and requirements.

Business-ready data means the data is not only error-free, but it will also support the business processes without causing an interruption. The master data fully supports the transactional data and all of the data supports the business processes across the enterprise.

The amount of data across an enterprise is staggering. Given the volume of data, understanding what data is relevant to the business is required. For example, the System for Award Management contains hundreds of thousands of vendors. Understanding which vendors are actually relevant to the business, which vendors are being used in the contract writing system, and which vendors have open obligations defines the vendors that are actually relevant to the business. One would analyze this data to identify exactly what vendors have open transactions or have had transactions recently, and use this to prioritize the data cleansing effort. Letting the data speak for itself identifies where the issues are and their scope, and allows the business to focus its efforts.

How to Let the Data Speak for Itself

To fully understand your data you must analyze all *business relevant* data. The key here is to first understand what is relevant. This needs to be based on facts. Analyze the data and develop fact-based analytics to answer these questions:

- What are the retention requirements?
- What data has open transactions (accounts payable, purchase orders, etc.)?
- Of that population, what configuration is being used?
- How often is new data created?

Once you have identified what is relevant to the business, you can analyze how the data is being used. Use the following questions for data usage insight:

- Is the business storing the same data in different places for different uses?
- Are text fields being used inappropriately – is the description of an award accurate and appropriate?
- Is the data being maintained correctly – are awards set to close at the appropriate time or left open?
- Is the data being used correctly – are funds inappropriately set up for labor costs?

Then, rinse and repeat; letting the data speak for itself is an iterative process. How the data is being used drives its relevance, and usage may evolve over time. This relevance exercise may also identify data that does not need to be collected anymore because it is not being used.

Authors & Affiliations

Christopher Babcock, PwC

Herschel Chandler, Information Unlimited, Inc.

Sharon Kuck, Dun & Bradstreet

Jason Ludwig, BackOffice Associates, LLC

Darla Marburger, Claraview