Do You Know Your Data? Does the Rest of the Company? Nothing Good is Built Without a Blueprint.

Eva M. Larson
Zebra Technologies Corporation

So what is data? It is a collection of facts from which conclusions may be drawn. It is neutral, being neither valid or invalid, nor accurate or inaccurate until it is strategically organized for use in analysis, reasoning and decision making.

Data is a valuable, competitive corporate asset and the quality of that data is crucial to the success of every company. We make important decisions using data. It helps identify where a company has been. It helps identify how a company is doing today and tomorrow and the next. It helps measure the progress of decisions made and it provides the ability to predict and forecast.

Today's technology has the ability to bring siloed department data together into one giant corporate sandbox where the data elements can be shared across the entire organization. The company's goal is a single source for master data, but how can this be accomplished when most business areas are used to working independently? Who owns the data? What type of controls and standards are required and how do you measure your successes?

The management of data is not a product it is a strategy. Data Management and Data Quality are not new concepts, just old ones that have been forgotten along the technology wave. Somehow over the years we accepted the falsehood (that I don't recall was ever formally verbalized) that technology was going to solve all of our problems. It can't, and it never will, not even if we end up creating a machine that can think, learn, and act like a human. There will always be the need for a beginning, a source, an original point of entry. There will be a someone, not a something, who will be responsible for performing the hard tasks of determining what type of data and processes are required to create and achieve the identified outcome.

Think about it. Data is everywhere and it is travelling at the speed of light these days, but it takes effort using people, processes and technology to harness the power that data can provide. And just like putting a puzzle together, the specific pieces must fit exactly right to crate the picture the producer intended for the consumer to experience. Without data management to correct data errors, products will be shipped to the wrong address, bills will be sent to the wrong person, and technology will continue to move a company faster to the costly and time consuming task of manual error correction procedures. Remember the term GIGO - garbage in, garbage out? This is a famous computer axiom meaning that if invalid data is entered into a system, the resulting output will also be invalid. As I mentioned previously, this concept is not new, just forgotten.

Look around you. Everything is designed with a purpose in mind, and, for most everything we understand what that purpose is based on the information provided to us. Labels on food products and clothing allow us to identify what the product is made of, who made it, where it was made, how to use it, and how to clean and care for it. Instruction manuals describe a product's functionality and how it is used. All of this information started somewhere. It started as individual bits of data, which once known and understood, were strategically architected by people using specific processes and in some cases technology (but not in all) for the purpose of a predetermined outcome. Typically the purpose is to make the lives of consumers easier. As consumers, we are where the original idea came from, and all the efforts taken to improve the customers' experience.

This paper should provide the essential details involved in achieving a synchronized, consistent view of an organization's core business entities. It assumes the reader has a basic understanding of the definitions, terms, and concepts surrounding data governance, data quality and metadata repositories.

THE PATH OF PROGRESS: FUTURE STATE

To build a blueprint for any project you need to know what the end state should look like. Determine the goals and objectives of the project and identify what data changes are necessary to meet that vision.

For Zebra Technologies it is a single source of truth for supplier, customer, and product master data with transactional data linkages attaching to multiple systems and applications. Find out what type of systems, application and software the company is planning to purchase and how all of it ties together. It means identifying what tools will be needed to accomplish this such as: data integration, data synchronization, data governance, and data quality. It also means following an enterprise wide sets of definitions, standards and processes. It is a place where everything required to be known about a customer is captured and used to promote the goal of enhancing the customer's experience and where all of the business areas are able to successfully use a shared set of high quality data.

Sounds simple on paper right? Without knowing the data it would be impossible. With knowing the data it is hard work but can be accomplished with a strategy in place.

THE PATH OF PROGRESS: CURRENT STATE

In order to get to where you want to be you need to know where you are starting from. Just like using MapQuest, you must know the starting and ending points in order to determine the best route. We have determined what the end state will look like now let's identify the current state.

INVENTORY

What exactly does *knowing* data mean? According to Webster's Unabridged Dictionary the word know means "to perceive with certainty; to understand clearly; to be sure of or well informed about; to have a clear and certain perception; to have knowledge". So knowing data, within the confines of our corporate world, means having a clear understanding of the data elements used by the company and how that data is used.

Data Architecture is the ability to design data to meet a future need. Its purpose is to describe how data is defined, processed, stored, and utilized. Let me begin our journey by stating that this process is not for the 'faint of heart'. It is a difficult and time consuming job but one that is necessary if we are to achieve our end goal of continuous and consistently synchronized, corporate quality data. Don't let anyone tell you that knowing data is not important, for without that knowledge, it is impossible to make wise and educated decisions. Without it, we would have to resort to the flip of the coin. Without it, we would have no one to market to, no one to send a bill or payment to, and no way to get directions when starting at point A and needing to get to point B.

The goal of every project is to get from where you are now to where you want to be. In order to achieve this, one must identify the data in its current state. Unfortunately, most companies are already established and do not allow for the luxury of starting from scratch. The good news is data is already being captured somewhere, most likely in a number of systems. These systems can be mined and compiled in order to get the ball rolling toward understanding the data.

Begin by creating an inventory of all of the data elements from all source systems required in the project scope, including an attempt to match like entities to one another. Include in this list the name of the data field, name of the source system, and the table and field name within each system. For example,

Data Element	Source1 Table and Field	Source2 Table and Field

Company Name	TTCCOM010.T$NAMA	HZ_PARTIES.PARTY_NAME
Account Name	TTCCOM010.T$NAMA	HZ_CUST_ACCOUNTS.ACCOUNT_NAME
Account Number	TTCCOM010.T$CUNO	HZ_CUST_ACCOUNTS.ACCOUNT_NUMBER

This information can also be incorporated into the data mapping and a corporate metadata registry.

Does this seem overwhelming? Can this really be accomplished when, for example, Oracle eBusiness Suite has over 20,000 tables with over 600,000 data fields? Yes it can, provided you start small and build incrementally. Start by identifying foundational data elements only. Foundational elements are often repeated throughout multiple systems and tables. For these types of data elements the purpose is to also leverage data for multiple uses. Examples of foundational elements may include: address, first name, or creation date.

Once the list has been narrowed down to a manageable number of elements, they should be grouped into different subject areas such as: customer master, product master or supplier master.

PROFILING

All that is known at this point are the names of the data elements and the systems in which each resides. Profiling is the task of identifying the current state of the data within each field. This is where a query tool is useful. Run queries against each of the fields as a fact gathering mission to identify if a field is being used and, what distinct values are in that field. Find out how often each value is used and whether or not there are nulls (a clue the field is not mandatory). In some cases, if you are lucky, mapping documents may even exist that can be reviewed for a wealth of information concerning the data requirements.

By adding this information to your inventory patterns will emerge. Similarities and differences between sources can be identified, and whether or not lists of values have been used vs. free form text. Note any clues that may identify if an element or value is no longer used. Summarize the facts gathered.

Identify questions concerning the data profiled. Even if you think you know the answer still write the question down. Remember every department has their own definitions of fields and values and different uses surrounding the data entered and consumed. When in doubt always let the business decide. After all they are the ones using the data.

Besides identifying what is in an individual field you will also attempt to identify any relationships between fields such as: country and state or province or person title to person name.

Document everything. You will need it later.

INTERVIEWING

Now that you have a foundational knowledge of the data, it is time to identify the details surrounding that data, the data values themselves, and the processes used for the production and consumption of that data. This can be done by interviewing the subject matter experts (SME) within each business area. This process may sound similar to the role of a Business Systems Analyst (BSA), but the goal is much different. A BSA’s focus is on determining how to get the data from one source into another and automating any of the manual processes along the way (which you will need to know as well). A Data Architects responsibility is to know data information or characteristics about the data (known as Metadata) with the goal of optimizing that data through the use of standards and rules. Even though both groups rely on interviews, the outcomes will be different. Based on the information that you gathered during the inventory task, ask questions, such as where does the data come from, what is it used for, who enters the data and why. Record the definitions of each data element and each data value. Do they have their own business processes that are followed, and if so, what are they? Don’t be afraid to ask for clarification. For example, a field called term might mean payment term to the Finance team but delivery term to the Receiving team. A field value of 30N in one department may be the same as 30 NET in another, both describing a term of Net 30 days. Don’t be surprised that each business area will have different answers to all of your questions. All of these differences (and similarities) must be documented.

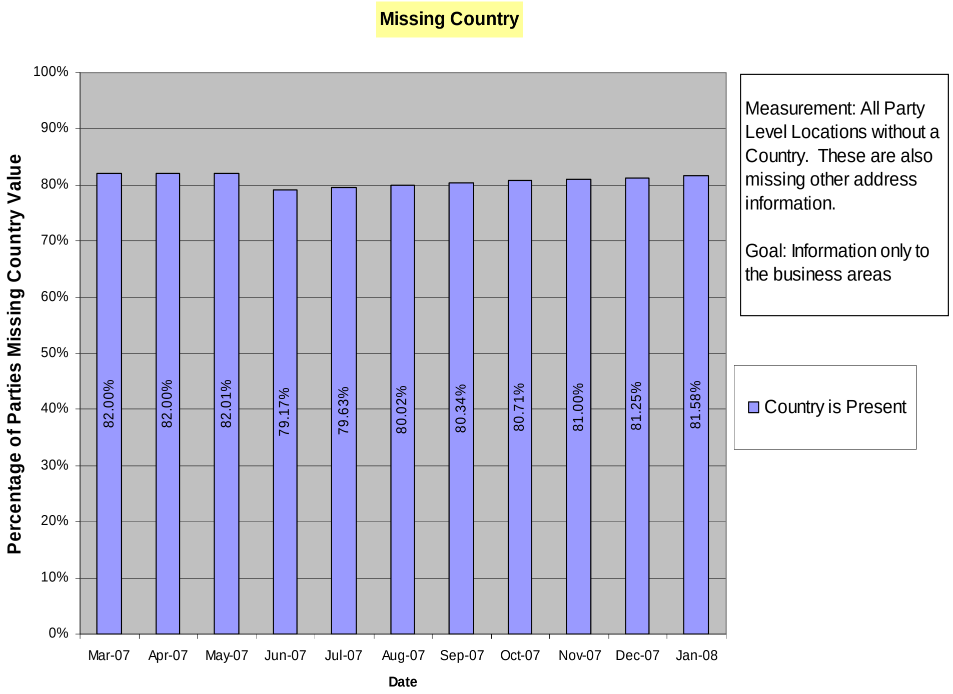
As you can see, this is time consuming but also most beneficial since all of the information will help determine the canonical version of definitions, standards and rules surrounding data entry (production) and data usage (consumption) for the entire company.

BASELINE METRICS

The best way to measure how far you have come after a project's implementation is through a series of graphical baseline metrics. This will allow the business areas and corporate executives a view into what the data looked like in its original source. These metrics will not only help provide proof of the necessity for data management and the improvement of the quality of that data but will also identify the magnitude of the tasks required to clean it all up.

Use the information gathered in the profiling process to create a number of metrics making sure they align with the business objectives of the project. Most important: Completeness, Consistent, Valid and Accurate. The term single source of truth does not only reflect a single view of an organization but the standardization of all the data elements within a system. The data needs to be reliable.

A picture paints a thousand words so the saying goes. The best way to present current state to management is through a series of graphs. Also, the metrics need to be repeatable so that after the project is in production, the same set of graphics can be created to show the successful reconstruction of that data. Though simple in nature, the graph below identifying the presence of a country code shows the business users that country code is not mandatory.



THE PATH OF PROGRESS: DATA BLUEPRINT

Using the information gathered during the current state review we can now begin selecting the data elements that will represent the customer, supplier or item master and begin piecing together a common set of standards, rules, controls and definitions for each master data element. This is a necessity because without it you cannot communicate effectively if every application is using a different set of terms and definitions. The primary requirement is a methodical, precise and meticulous identification of each data element that is communicated in such a way that there will be no ambiguities surrounding the entry, meaning, and use of that element. The ultimate final decisions however are not made by IT but by the business areas themselves. IT's responsibility is to gather the facts, provide recommendations, and facilitate the final decision making. They are the gate keepers and tool providers while those entering and using the data are the experts and have ultimate responsibility for the quality of that data.

This information will need to be used throughout the entire project and beyond. The decisions made above need to be applied to data being converted, interfaced, manually entered, transmitted via EDI or webforms, including 3rd party vendors. I read a funny joke while perusing Wikipedia. I was searching for Data Architect and here is what I found "What's the difference between a Data Architect and a terrorist? You can negotiate with a terrorist!" Quality must be the number one priority and non-negotiable to keep the data in sync with the standards set. To achieve this, the information must be communicated throughout the enterprise in both verbal and written form. This means creating and maintaining a metadata dictionary and forming data stewardship teams.

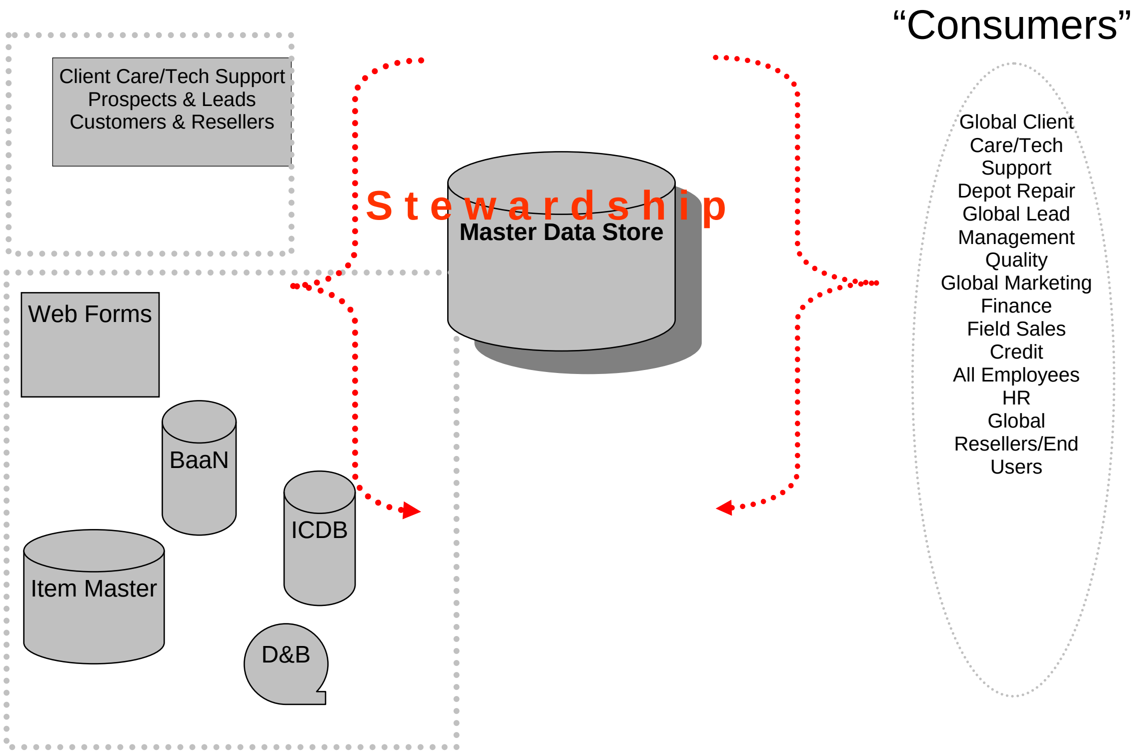
METADATA DICTIONARY

A good written communication tool is a Metadata Dictionary and one should be created specific to each subject area. I am not talking about a typical IT data dictionary but one that will communicate information to both the business areas and IT. It is also the place where standards and processes are written so that when issues come up or changes are required the foundation or starting points are known. The goal of a dictionary is to bring order and control by identifying consistent definitions of data and procedures so that data remains accurate over time. The end result will also be reliability and confidence that the entry of information is consistent across all areas of input.

Each data element should have its own page and the information within that page should communicate the details surrounding that particular data element. At the very least the information identified for each data element should include:

- Table of Contents page
- Data Element Name
- List of synonym names for each data element that are used by the business areas
- Data Element Definition
- Table and field names from all systems
- List of Contributors (data entry)
- List of Consumers (reporting)
- Key Uses
- Data Flow diagrams or link to process diagrams
- Data Entry Rules
- Field value standards including lists of values
- Processing rules
- Length of field
- Field Type
- Data Steward/Business Area Expert responsible for data element
- Survivorship rules for data transfer and Deduplication

Besides identifying the data elements it would be beneficial to include a picture of how the data is processed. Below is a very simplistic design meant to provide an idea for building one within your own company.



In order for the dictionary to be easily accessible I recommend posting it in a place where all employees have the ability to access it. Highest visibility would be an intranet site or viewable as online Help within the system itself.

The information in this dictionary is not static. Companies are constantly changing how they do business and so their data requirements will change as well. It will need to be continuously maintained to incorporate all changes related to the data. To ensure this a data stewardship team should be in place.

DATA STEWARDSHIP

A good verbal communication strategy is to utilize data governance and a data stewardship team for each subject area. Besides data, business processes cover everything an organization does with the data. Automating these is important however, the business areas have to be in agreement concerning how data elements, applicable to multiple areas, should be defined, and used. Since master data elements are used by multiple business areas it is highly recommended that a data governance executive and data stewardship team be created.

Data stewards are trustees of the data, not owners, and are responsible for establishing strategy, objectives, and policies that drive data quality improvements. Each steward is responsible for managing the data created by their respective business areas, identifying definitions, quality standards and security. They must be given the authority to manage the data so that it reflects the desired business outcomes defined within the metadata dictionary. These teams should consist of subject matter experts from each major business area and worldwide if a company does business internationally. A good source for identifying these members might be the same people identified during the interview process. Each representative should have the ability to drive change. They need to be able to operate at both a strategic and tactical level. A major challenge is that the business areas have widely differing perspectives concerning the data so they will need to bring together diversity and unity while supporting each of the individual members.

Because of the differing perspectives, a data governance executive sponsor (for each subject area) is a necessity and should be a person from one of the business areas. This executive assists the team in resolving conflicts, driving the standards and requirements across all business areas and implementing performance measurements to ensure that the quality processes are enforced.

Participation should be mandatory since these teams will have the ultimate responsibility for identifying and enforcing data standards throughout the enterprise. These teams are not just for the current project, but for ensuring that quality is maintained for all on-going business processes. Team members must also be active participants in taking responsibility for quality of the corporate data and educating all employees within their business areas.

Specific goals should be set by the teams with the focus on data quality improvements. These need to be well defined and measureable. Guiding principles should be written so that the team can stay focused on their purpose and goals. Regular meetings should be established along with rules and guidelines concerning how those meetings will be conducted.

METRICS

Since data has significant impact on a company's most strategic business initiatives, the quality of that data should be measured on a regular basis. According to Stephen Hawking in his 1998 book A Brief History of Time "…entropy…measures the degree of disorder of a system. It is a matter of common experience that disorder will tend to increase if things are left to themselves…one can create order out of disorder…but that requires expenditure of effort." The most strategic way to keep data clean is to position the quality standards and processes closest to the point of entry however it still needs to be reviewed in order to verify its worth. Every aspect possible should be measured (a good place to start identifying what to measure is the information housed in the metadata dictionary) and communicated to the entire corporation in a graphical format. Metrics will not only help ensure that business rules and standards are being followed but they will also help identify areas of improvement.

Metrics can be separated into specific groups such as volume metrics, completeness checks, validity, accuracy between related data elements and interface comparisons.

For example, volume metrics could be used to identify how many organizations or contacts, by responsibility, have been entered since the system was first moved into production. These types of metrics can show growth areas as well when captured on a monthly basis.

Completeness checks identify where data elements are missing. Again it can be identified by responsibility, region or whatever makes sense to your organization.

Valid and accurate checks identify where fields may be in adherence to a static list of values, or document where new or old values exist. These also can be used to pair up key relationships to verify that fields that should relate to each other actually do so. For example, a country relationship to province exists where the country is Canada. In this example, the provinces can only be chosen from a set of 13 specific values.

Interface comparisons are important when data is being transported to and from a variety of applications and helps measure the consistency of that data. It is always a good idea to perform regular sanity checks to verify that the same data within multiple systems does indeed match up.

Detailed metrics of non-compliance should be reviewed in order to identify the source(s) of the problem. First fix the non-compliance issue and then fix the data history. It is recommended that the department that created the non-compliance problem should be responsible for fixing the data history. Otherwise, the problems may never get resolved.

"TO INFINITY AND BEYOND"

Ah the end…but there never really is an end to data management and data quality. Quality should always be a priority for any initiative and even after most projects have been completed, data quality is an ongoing process that will outlast any other IT initiative.

Now that you have your blueprint in place, with business areas and IT educated on master data and the information surrounding each data element, a process needs to be identified for any updates to existing entities and processes or for anything new. The secret to future success is establishing the rule that anything directly or indirectly affecting any master data element must be presented to the data governance/stewardship teams. The purpose of this is to maintain data quality and standards for each and every master data element.

Remember what I mentioned at the beginning? The management of data is not a product, it is a strategy. It needs to be designed to continue working long after a project has been completed, or employees have left their positions. The truth is data and the quality of that data is everyone's business. It takes teamwork like the Three Musketeers "All for one and one for all" or like the Verizon commercial where the entire company is present to support the customer. It takes team effort to produce data…and if you are going to go through all the effort of capturing data it might as well be done right the first time by capturing quality data. Then we can all spend more time on strategy instead of on clean up. Focus can be given to driving enterprise changes rather than expending all resources on production maintenance.