CHAPTER 4

Data Quality Assurance

The previous chapters define accurate data. They talk about the importance of data and in particular the importance of accurate data. They describe how complex the topic really is. You cannot get to accurate data easily. They show that data can go wrong in a lot of different places. They show that you can identify much but not all inaccurate data and that you can fix only a small part of what you find.

Showing improvements in the accuracy of data can be done in the short term with a respectable payoff. However, getting your databases to very low levels of inaccuracies and keeping them there is a long-term process.

Data accuracy problems can occur anywhere in the sea of data residing in corporate information systems. If not controlled, in all probability that data will become inaccurate enough to cause high costs to the corporation. Data accuracy problems can occur at many points in the life cycle and journeys of the data. To control accuracy, you must control it at many different points. Data can become inaccurate due to processes performed by many people in the corporation. Controlling accuracy is not a task for a small, isolated group but a wide-reaching activity for many people.

Data accuracy cannot be "fixed" one time and then left alone. It will revert back to poor quality quickly if not controlled continuously. Data quality assurance needs to be ongoing. It will intensify over time as the practitioners become more educated and experienced in performing the tasks necessary to get to and maintain high levels of data accuracy.

This chapter outlines the basic elements of a data quality assurance program. It focuses on data accuracy, a single dimension of data and information quality. This is not to mean that the other dimensions should not also be

addressed. However, data accuracy is the most important dimension, and controlling that must come first.

4.1 Goals of a Data Quality Assurance Program

A data quality assurance program is an explicit combination of organization, methodologies, and activities that exist for the purpose of reaching and maintaining high levels of data quality. The term *assurance* puts it in the same category as other functions corporations are used to funding and maintaining. Quality assurance, quality control, inspection, and audit are terms applied to other activities that exist for the purpose of maintaining some aspect of the corporation's activities or products at a high level of excellence. Data quality assurance should take place alongside these others, with the same expectations.

Just as we demand high quality in our manufactured products, in our financial reports, in our information systems infrastructure, and in other aspects of our business, we should demand it from our data.

The goal of a data quality assurance program is to reach high levels of data accuracy within the critical data stores of the corporation and then keep them there. It must encompass all existing, important databases and, more importantly, be a part of every project that creates new data stores or that migrates, replicates, or integrates existing data stores. It must address not only the accuracy of data when initially collected but accuracy decay, accurate access and transformation of that data, and accurate interpretation of the data for users. Its mission is threefold: improve, prevent, monitor.

Improvement assumes that the current state of data quality is not where you want it to be. Much of the work is to investigate current databases and information processes to find and fix existing problems. This effort alone can take several years for a corporation that has not been investing in data quality assurance.

Prevention means that the group should help development and user departments in building data checkers, better data capture processes, better screen designs, and better policies to prevent data quality problems from being introduced into information systems. The data quality assurance team should engage with projects that build new systems, merge systems, extract data from new applications, and build integration transaction systems over older systems to ensure that good data is not turned into bad data and that the best practices available are used in designing human interfaces.

Monitoring means that changes brought about through data quality assurance activities need to be monitored to determine if they are effective. Monitoring also includes periodic auditing of databases to ensure that new problems are not appearing.

4.2 Structure of a Data Quality Assurance Program

Creating a data quality assurance program and determining how resources are to be applied needs to be done with careful thought. The first decision is how to organize the group. The activities of the group need to be spelled out. Properly skilled staff members must be assigned. They then need to be equipped with adequate tools and training.

Data Quality Assurance Department

There should be a data quality assurance department. This should be organized so that the members are fully dedicated to the task of improving and maintaining higher levels of data quality. It should not have members who are part-time. Staff members assigned to this function need to become experts in the concepts and tools used to identify and correct quality problems. This will make them a unique discipline within the corporation. Figure 4.1 is a relational chart of the components of a data quality assurance group.

The group needs to have members who are expert data analysts. Analyzing data is an important function of the group. Schooling in database architecture and analytical techniques is a must to get the maximum value from these activities. It should also have staff members who are experienced business analysts. So much of what we call quality deals with user requirements and business interpretation of data that this side of the data cannot be ignored.

The data quality assurance group needs to work with many other people in the corporation. It needs to interact with all of the data management professionals, such as database administrators, data architects, repository owners, application developers, and system designers. They also need to spend a great deal of time with key members of the user community, such as business

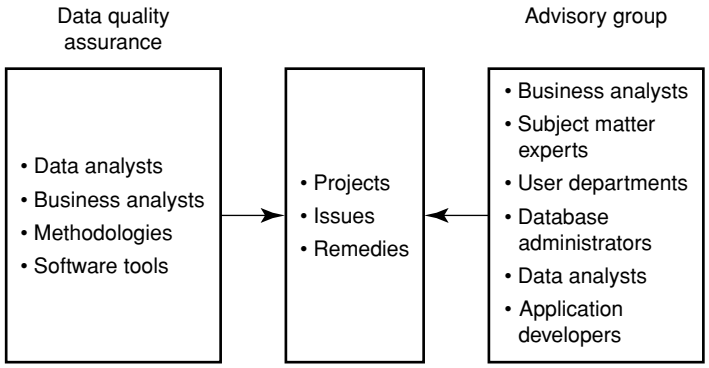

FIGURE 4.1 Components of a data quality assurance group.

analysts, managers of departments, and web designers. This means that they need to have excellent working relationships with their customers.

THERE *is a strong parallel between the emergence of data quality assurance to the improvements made in software development in the 1970s and 1980s. Software development teams back then consisted mostly of programmers. They wrote the code, tested the product, and wrote the user manuals. This was the common practice found in the best of software development groups.*

In my first job at IBM I designed, developed the code, tested, wrote user documents, and provided customer support of a software product (Apparel Business Control System). It was a one-person project. Although the product had high quality and good customer acceptance, I believe it would have gone better and been a better product if I had access to professional writers and software quality assurance people.

In response to the continual problems of poorly tested products and very poor user manuals, companies started dedicating some of the programmers to ensuring the quality of code (testing) and began to hire professional technical writers. There was an immediate improvement in both the code and user manuals. As time went on, these two areas became established disciplines. Software development companies specialized in building tools for these disciplines; colleges offered classes and tracks for these disciplines.

The programmers that tested were no different from those that wrote the code in the beginning. They made huge improvements only because they were dedicated to testing, worked with the programmers throughout the entire project, and brought another view to the use of the code. In time, they became even better as they developed very effective methodologies and tools for testing. Testing became a unique technology in its own right.

The cost of these programs is clearly zero. Every serious development group today separates code quality assurance from code development. Projects finish earlier, with higher-quality results. The projects spend less money (much less money) and use up less time (much less time) than they would if programmers were still doing the testing.

Data quality is emerging as a major topic 20 years later. The same evolution is happening. Making data quality the responsibility of the data management staff who design, build, and maintain our systems means that they do not become experts in the methodologies and tools available, do not have the independence to prioritize their work, and do not focus on the single task of ensuring high-quality data. Data quality assurance must be the full-time task of dedicated professionals to be effective.

One way to achieve a high level of cooperation is to have an advisory group that meets periodically to help establish priorities, schedules, and interactions with the various groups. This group should have membership from all of the relevant organizations. It should build and maintain an inventory of quality assurance projects that are worth doing, keep this list prioritized, and assign work from it. The advisory group can be very helpful in assessing the impact of quality problems as well as the impact of corrective measures that are subsequently implemented.

Data Quality Assurance Methods

Figure 4.2 shows three components a data quality assurance program can build around. The first component is the quality dimensions that need to be addressed. The second is the methodology for executing activities, and the last is the three ways the group can get involved in activities.

The figure highlights the top line of each component to show where a concentration on data accuracy lies. Data accuracy is clearly the most important dimension of quality. The best way to address accuracy is through an inside-out methodology, discussed later in the book. This methodology depends heavily on analysis of data through a process called data profiling. The last part of this book is devoted to explaining data profiling. Improving accuracy can be done through any of the activities shown. However, the one that will return the most benefit is generally the one shown: project services.

Any data quality assurance function needs to address all of the dimensions of quality. The first two, data accuracy and completeness, focus on data stored in corporate databases. The other dimensions focus on the user community and how they interpret and use data.

The methods for addressing data quality vary as shown in Figure 4.3. Both of these methodologies have a goal of identifying data quality issues. An

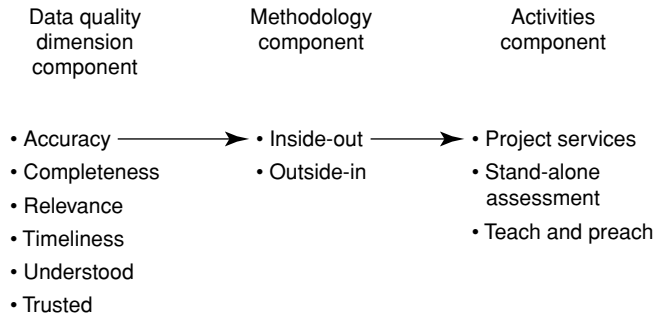

FIGURE 4.2 Components of a data quality assurance program.

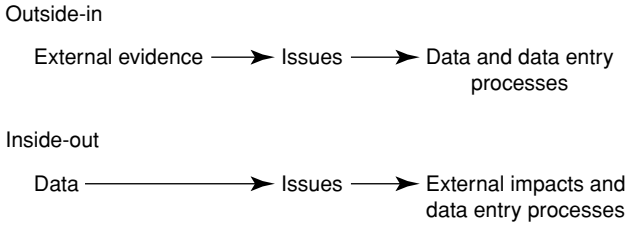

FIGURE 4.3 Methodology comparisons.

issue is a problem that has surfaced, that is clearly defined, and that either is costing the corporation something valuable (such as money, time, or customers) or has the potential of costing the corporation something valuable. Issues are actionable items: they result in activities that change the data quality of one or more databases. Once identified, issues are managed through an issues management process to determine value, remedies, resolution, and monitoring of results. The process of issue management is discussed more fully in the next chapter.

INSIDE-OUT METHOD

The inside-out method starts with analyzing the data. A rigorous examination using data profiling technology is performed over an existing database. Data inaccuracies are produced from the process that are then analyzed together to generate a set of data issues for subsequent resolution.

The analysis should be done by a highly qualified data analyst who understands the structure of the data. The methodology starts with a complete and correct set of rules that define data accuracy for the data. This is metadata. It consists of descriptions of the data elements, values permitted in them, how they relate to one another in data structures, and specific data rules that describe value correlation conditions that should always be true within the data. All of these categories are discussed at length in later chapters.

Of course, such a rigorous rule set for any operational database does not exist. The metadata that is available is generally incomplete and most likely inaccurate. The data profiling process described in later chapters is a process that completes and corrects the metadata, along with using it to find evidence of inaccurate data. This intertwined process has a very valuable by-product: accurate and complete metadata.

The process of determining the correct metadata inevitably involves conferring with business analysts and end users. The data analyst will detect a behavior in the data and require consultation to determine why it is so. This often leads to modifications to the metadata. These consultations are

always productive because the question is always backed up by information from the data.

The data analyst should identify who in the user community will be the most valuable in consulting on issue identification and form a small, dynamic working group with them. In the end, they should always agree on what the final metadata is, and agree on the inaccurate data facts derived from the comparison with the actual data.

The inaccurate data evidence produced is a collection of facts. It may be explicit cases of wrong or missing values, or it may identify rules that fail without being able to say what values are wrong. For example, one fact may be that 30% of purchase order records do not have a supplier ID. Another may be that the employee birth date field has values that are invalid: too long ago or too recent. Another might be that the percent of the color BLUE in a database is too large. In this case, the analyst does not know which instances are correct and which are wrong; only that some of them must be wrong.

The facts are aggregated into issues. Some facts are issues by themselves. For example, the supplier ID problem may be the basis for a single issue. Others are aggregated into a larger issue. An example is that customer demographic fields in a marketing database contain numerous errors in all fields, possibly indicating a general problem with form design.

OUTSIDE-IN METHOD

This method looks for issues in the business, not the data. It identifies facts that suggest that data quality problems are having an impact on the business. It looks for rework, returned merchandise, customer complaints, lost customers, delays in getting information products completed, high amounts of work required to get information products produced, and so on. Interviews are done with users to determine their level of trust in the accuracy of data coming from the information systems and their level of satisfaction with getting everything they need. It may also include looking for decisions made by the corporation that turned out to be wrong decisions.

These facts are then examined to determine the degree of culpability attributable to defects in the data. The data is then examined to determine if it has inaccuracies that contribute to problems, and to determine the scope of the contribution. This examination is generally pointed at the specific problem. It is generally not a thorough data profiling exercise, although it could be expanded to that if the evidence indicates a widespread quality problem with the data.

This approach is generally the work of the data quality assurance team member with skills as a business analyst. It involves heavy participation on the

part of outside people. It also requires conference sessions with user community experts. The result is a collection of data issues that are then tracked on the same path as those from the inside-out methodology.

COMPARISON OF METHODS

Neither approach is superior to the other: they both bring value to the process. However, they do not get to the same end point. Data quality assurance groups should use both methodologies as applicable.

Inside-out is generally easier to accomplish and uses less people time. A single analyst can analyze a great deal of data in a short time. The data quality assurance group can accomplish a great deal with this approach with the staff within their own department. The outside-in approach requires spending a lot of time interviewing people in other departments.

The inside-out approach is nondisruptive. You just get a copy of the data you want to analyze and do it offline. The outside-in approach requires scheduling time for others, thus interrupting their regular activities.

The inside-out approach will catch many problems the outside-in approach does not catch. For an outside-in approach to catch a problem, it must manifest itself in some external behavior, and that behavior must be recognizable as being not good.

An example of a hidden problem is a case in which missing supplier ID numbers on purchase orders causes a company not to get maximum discounts they were entitled to from suppliers. The purchase order volumes were summarized by supplier ID and, because the field was missing on 30% of the records, the amounts were low. The company was losing millions of dollars every year because of this and was completely unaware that it was happening. The inside-out approach catches this; the outside-in approach does not.

Another type of problem are those inaccuracies that have the potential for a problem but for which the problem has not yet occurred. An example of this is where an HR database failed to capture government classification group information on employees accurately. Many minority employees were not classified as minorities, nor were handicapped employees all being identified as handicapped. No problem may have surfaced yet. However, the potential for being denied contracts in the future because of these inaccuracies is waiting to happen. Inside-out analysis will catch this; outside-in will not.

The opposite is also true. The inside-out approach will not catch problems where the data is inaccurate but valid. The data can pass all metadata tests and still be wrong. This can happen either because the rule set is incomplete or because the data hides underneath all of the rules. An example is getting the part number wrong on orders. The wrong merchandise is shipped. An analysis of the data will not reveal inaccurate data because all of the part num-

bers are valid numbers. The outside-in approach catches these problems better. (The inside-out approach may catch this if the analysis finds the percentage of orders returned to be higher than an acceptable threshold. This is possible if a data rule or value test has been formulated. These topics are covered in Chapters 11 and 12).

There is another class of problems not detectable by either approach. The data is valid but wrong and also produces insufficient external evidence to raise a flag. Although these generally are of little concern to a corporation, they have the potential to be costly in the future if not detected. A data quality assurance program built exclusively using only one approach is generally going to miss some important issues.

Data Quality Assurance Activities

The data quality assurance team must decide how it will engage the corporation to bring about improvements and return value for their efforts. The group should set an explicit set of guidelines for what activities they engage in and the criteria for deciding one over another. This is best done with the advisory group.

There are three primary roles the group can adopt. This is shown as the last column in Figure 4.2. One of them, project services, involves working directly with other departments on projects. Another, stand-alone assessments, involves performing assessments entirely within the data quality assurance group. Both of these involve performing extensive analysis of data and creating and resolving issues. The other activity, teach and preach, involves educating and encouraging employees in other groups to perform data auditing functions and to employ best practices in designing and implementing new systems.

PROJECT SERVICES

The vast majority of projects being pursued by the IT organization involve repurposing an existing database. It is rare these days to see a truly new application being developed that does not draw from data that has already been collected in an existing application. Examples of projects that involve working with existing data stores are

- data migration to new applications (generally packaged applications)
- consolidation of databases as a result of mergers and acquisitions
- consolidation of databases to eliminate departmental versions of applications

- replication of data into data warehouses, data marts, or operational data stores
- building a CRM system
- application integration that connects two or more applications
- application integration that connects an older database to the Internet

There is a real danger in all of these applications of introducing errors through mistakes made due to a misunderstanding of the data. There is also a real danger in the data from the original systems not being of sufficient quality to meet the demands of the new use of the data. Both of these are classical concerns that if not addressed will certainly cause great difficulty in completing the projects, as well as unhappiness with the outcome.

The data quality assurance team can provide an invaluable service to these projects by profiling the data. By doing this they provide two valuable outputs: an accurate and complete metadata description of the data and an inventory of data quality problems uncovered in the process.

The metadata repository produced should be used to match target system requirements against the content and structure of the source systems. It is also the perfect input to developing processes for extraction, transformation, cleansing, and loading processes.

The data quality assurance team can use the inaccuracy facts to determine either whether the data is strong enough to satisfy the intended use or whether there is a need to establish new projects from the issues to drive improvements in the source systems. Of course, this applies to cases in which the source databases continue to live past the project, as is the case for replication and integration projects.

The data quality assurance team can also provide advice and oversight in the design of target database structures, as well as processes for collecting or updating data. They also have a good opportunity to get data checking and monitoring functions embedded in the new systems to help prevent future quality problems.

Why should the data quality assurance team perform these tasks, as opposed to the project teams? The answer is that the data quality assurance team are experts in data quality technologies. They are experienced in data profiling, investigation of issues, and fabrication of data quality problem remedies.

One of the most valuable outputs of data profiling at the beginning of a project is to learn that the project cannot achieve its goals because of the condition of the source data. When this happens, the project team can then make decisions about changing target design, changing target expectations, making improvements to data sources, or scrapping the project outright. This is the

perfect place to make these decisions: before most of the project money has been spent and before most of the development work has been done.

Projects that do not perform a thorough review of the source data generally do not discover the match between the data and the project requirements until after much time and money has been spent. It is generally very expensive to repair the damage that has already been done and impossible to recoup the money spent and the valuable time lost.

STAND-ALONE ASSESSMENTS

A stand-alone assessment is a project organized for the purpose of determining the health of an existing database. The database is chosen because of suspicions or evidence about problems coming from the use of the data, or simply because it is an important data source for the corporation.

The data quality assurance team will generally execute the entire project. Using the inside-out method, they will profile the data, collect quality facts, produce issues, and then follow the issues through to remedies.

The advantage of assessment projects is that they do not require as much interaction with other project teams and can be scheduled without concern for other plans in IT. Of course, it makes no sense to schedule an assessment of a database that is about to get a facelift as a result of another project.

An assessment can be quite disruptive to other departments, even if no change activity is under way for the data source. Time from them will be needed to develop perfect understanding of the metadata and to interpret facts that come out of profiling. If remedies are needed, negotiations with IT and users will be needed to get them designed and implemented. It may also be quite disturbing to people to find out that they have been using flawed data for a long time without knowing it. The data quality assurance team needs to involve the other departments in the planning phase and keep them involved throughout the process.

It is important not to appear as an outside hit team trying to do damage to the reputation of the operational organizations. Involving them makes them part of the solution.

TEACH AND PREACH

This function involves training information system staff members on the technology available for data quality assessment, the techniques and best practices available for building and maintaining systems, and how to develop quality requirements and use them to qualify data.

Few information systems professionals come out of college with training explicitly targeted to data quality. The principles are not difficult to under-

stand, nor are the disciplines difficult to use in daily practice. Educating them will improve all of the work they do.

The data quality assurance group should function as the experts in data quality. They should not keep this knowledge exclusively to themselves. The more they educate others in the corporation, the more likely the information systems will reach and stay at a high level of quality.

Preaching means that the data quality assurance department should encourage and insist that quality checkpoints be put into all projects. They should encourage upper management to be cognizant of the need for data quality activities. They should collect and advertise the value to the corporation realized from these activities.

The data quality assurance group should not depend exclusively on teaching and preaching. If that is all they do, the company will never develop the focused expertise needed to analyze the mountains of data and drive improvements.

4.3 Closing Remarks

If you want high data quality you must have highly accurate data. To get that you need to be proactive. You need a dedicated, focused group.

You need to focus on data accuracy. This means you need an organization that is dedicated to improving data accuracy. You also need trained staff members who consider the skills required to achieve and maintain data accuracy as career-building skills.

You need to use technology heavily. Achieving high levels of data accuracy requires looking at data and acting on what you see. You need to do a lot of data profiling. You need to have experienced staff members who can sniff out data issues.

You need to treat information about your data as of equal or greater importance than the data itself. You must install and maintain a legitimate metadata repository and use it effectively.

You need to educate other corporate employees in the importance of data and in what they can do to improve the accuracy. This includes the following elements.

- Business users of data need to be sensitized to quality issues.
- Business analysts must become experts on data quality concepts and play an active role in data quality projects.
- Developers need to be taught best practices for database and application design to ensure improved data accuracy.

- Data administrators need to be taught the importance of accuracy and how they can help improve it.
- All employees who generate data need to be educated on the importance of data accuracy and be given regular feedback on the quality of data they generate.
- The executive team needs to understand the value of improved data accuracy and the impact it has on improved information quality.

You need to make quality assurance a part of all data projects. Data quality assurance activities need to be planned along with all of the other activities of the information systems department. Assisting a new project in achieving its data quality goals is of equal or higher value than conducting assessment projects in isolation. The more integrated data quality assurance is with the entire information system function, the more value is realized. And finally, everyone needs to work well together to accomplish the quality goals of the corporation.