The Informatica Data Quality Methodology

A Framework to Achieve Pervasive Data Quality
Through Enhanced Business-IT Collaboration
Table of Contents

Executive Summary ......................................................... 2

Meeting the Data Quality Challenge ................................. 3
  The Importance of Business-IT Collaboration ....................... 5
  Role-Based Tools for Enhanced Collaboration ....................... 5
Step 1: Profile the Data for Content, Structure, and Anomalies .... 7
Step 2: Establish Data Quality Metrics and Define Targets ......... 8
Step 3: Design and Implement Data Quality Business Rules ........ 9
Step 4: Build Data Quality Rules into Data Integration Processes .... 10
Step 5: Review Exceptions and Refine Rules .......................... 11
Step 6: Monitor Data Quality Versus Targets ......................... 12
Conclusion ................................................................. 13
Executive Summary

The three elements of any data quality initiative are people, processes, and technology. A structured, well-defined methodology is essential to orchestrating these three elements to derive the greatest payback from a data quality initiative.

While the value of a data quality methodology may seem self-evident, too many organizations approach data quality initiatives with ill-defined plans that introduce risks of confusion, overlooked details, redundant efforts, and subpar results.

A strategic and systematic methodology enables you to properly scope your data quality project, engage business and IT stakeholders with clearly defined roles and responsibilities, and equip them with the right technology and tools to tackle the data quality challenge.

This white paper examines the implications of poor data quality and introduces the Informatica® data quality methodology, a six-step framework that extends from initial profiling to continuous monitoring, toward the objective of making high-quality data pervasive throughout the enterprise.

It shows you how your business and IT users—business analysts, data stewards, and IT developers and administrators—can collaboratively use the Informatica data quality solution through each of the six steps to embed data quality across all data domains and applications throughout the extended enterprise.
Meeting the Data Quality Challenge

The performance of your business is tied directly to the quality and trustworthiness of its data. With high-quality data, your business is poised to operate at peak efficiency. High-quality data improves your competitive advantage and enhances your ability to:

- Acquire and retain customers
- Optimize sales and financials
- Run efficient supply chain and production processes
- Eliminate costly operational errors
- Make smart, timely business decisions
- Rapidly penetrate new markets

While most businesses recognize the theoretical importance of data quality, many wait until poor-quality data takes a bite out of operational efficiency and profitability before taking action. Consequences can range from customer service degradation, supply chain mistakes, and financial reporting errors to major operational failures that can cost millions of dollars a year.

Similarly, organizations often take an ad hoc approach to data quality—implementing quick fixes at a departmental or functional level that fail to comprehensively address data quality weaknesses across the enterprise and are ultimately short-sighted and unsustainable.

The costs are high. More than 140 companies surveyed by the analyst firm Gartner estimated they were losing an average of $8.2 million a year because of poor data quality. Losses of more than $20 million a year were cited by 22 percent of respondent organizations, and 4 percent put annual losses at more than $100 million.¹

"While losses of millions of dollars are significant, we believe these estimates understate the true financial impact on most organizations—the actual magnitude of the problem is typically far greater (by orders of magnitude) than is perceived by business and IT leaders," Gartner's report says.

To attack this problem, organizations need to invest in the people, processes, and technologies necessary to transform flawed data into trusted, actionable business information available to all stakeholders whenever and wherever they need it. The best data quality initiatives have these four characteristics:

- **Collaborative.** Business and IT share responsibility for data quality, with clearly defined roles and technology suited to the unique skills and perspectives of business analysts, data stewards, and IT developers and administrators.

- **Proactive.** Business and IT recognize that all organizations suffer some degree of poor data quality and proactively profile data to identify and correct problems before they materially impact business performance.

- **Reusable.** Data profiling and cleansing business rules can be reused across any number of applications to streamline and accelerate processes and help ensure high standards of quality.

- **Pervasive.** The data quality environment will extend to all stakeholders, data domains, projects, and applications regardless of where the data resides, whether on premise, with partners, or in the cloud.

For data quality to be most effective, it needs to be driven by a methodology that incorporates the characteristics defined above. Ideally, the methodology will be overseen and implemented by a data governance body, or it may be formalized in a center of excellence.

Informatica’s six-step methodology is designed to help guide data quality from the initial step of profiling to the ongoing discipline of continuous monitoring and optimization. Over nearly 10 years, the Informatica data quality methodology has evolved to become a mature and proven framework that has helped guide implementations in organizations around the world.

The methodology aligns with the Informatica data quality solution, which delivers the full range of data quality capabilities that your company needs to ensure that all its data is complete, consistent, accurate, and current. The solution consists of several packages optimized for specific uses: Informatica Data Quality™, Informatica Data Explorer™, and Informatica Identity Resolution™.

- **Informatica Data Explorer.** With role-based tools to promote collaboration between business and IT, this data profiling software discovers and analyzes the content, structure, and deficiencies of any type of data, in any source.

- **Informatica Data Quality.** The software executes cleansing, parsing, standardization, and matching processes and enables ongoing monitoring in visual scorecards or dashboards. As with Informatica Data Explorer, it features role-based tools to enable business and IT to work together.

- **Informatica Identity Resolution.** This software enables organizations to search and match identity data from more than 60 countries in batch and real time, across multiple enterprise or third-party applications.
The Importance of Business-IT Collaboration

A lack of collaboration between business and IT is a key reason why many data quality projects fail to live up to their potential. The two camps have traditionally relied on spreadsheets, documents, emails, and other tedious and imprecise mechanisms to communicate on data quality requirements.

Inevitably, it’s difficult for business analysts and data stewards to outline data quality business requirements in clear-cut terms that IT can understand. Misinterpretation, delays, high costs, and subpar results are common simply because business and IT are speaking two different languages, with no common framework. Critical details can be lost in translation.

Greater business-IT collaboration is increasingly recognized as essential to data quality and related data management initiatives. For instance, 64 percent of respondents to a survey by The Data Warehousing Institute reported that collaboration is an issue for data integration in their organizations.2

“More business people are getting their hands on data integration,” TDWI research senior manager Philip Russom wrote in TDWI’s What Works magazine. “Stewardship for data quality has set a successful precedent. This form of collaboration ensures that data integration truly supports the needs of the business.”

Role-Based Tools for Enhanced Collaboration

The Informatica data quality solution provides a foundation for collaboration between business and IT. It features role-based tools engineered to enable business analysts, data stewards, and IT developers and administrators to make the most of their unique skill sets and communicate with all stakeholders in the process.

These role-based tools present different views of the same data tailored to both business and IT. For instance, the IT developer sees technical versions of data and rules in a development environment. The business analyst sees a nontechnical rendering of the same data in a browser-based tool. Business and IT can work with identical data and rules, in terms that each understands, in a common environment that promotes joint ownership.

Shareable bookmarks and notes to communicate findings, requirements, results, and status enable team members to accelerate and streamline data quality processes, across multiple project groups, geographic locations, and time zones. Rules can be developed from these communications and viewed as part of the profiling results, greatly reducing the risk of misunderstandings about requirements.

---

2The Data Warehousing Institute, “Collaborative Data Integration,” TDWI What Works, August 2009.
Three role-based tools—Informatica Analyst, Informatica Developer, and Informatica Administrator—are common to both Informatica Data Explorer and Informatica Data Quality.

- **Informatica Analyst**: For Business Analysts and Data Stewards. By rendering data in semantic terms, the browser-based tool equips business analysts and data stewards with capabilities to profile data, create and analyze quality scorecards, manage exception records, develop and use rules, and collaborate with IT.

- **Informatica Developer**: For IT Developers. The Eclipse-based environment allows developers to discover, access, analyze, profile, and cleanse data, regardless of its location. Developers can model logical data objects, combine data quality rules with sophisticated transformation logic, and conduct mid-stream profiling to validate and debug logic as it’s developed.

- **Informatica Administrator**: For IT Administrators. This tool gives IT administrators centralized configuration and management capabilities. Administrators can monitor and manage security, user access, data services, and grid and high-availability configurations.

With an understanding of the components in the Informatica data quality solution, we can examine the six steps of the Informatica data quality methodology and how stakeholders can use the technology in each step. Figure 1 illustrates these six steps.

Figure 1. The Informatica data quality methodology extends from an initial profiling phase to ongoing monitoring and optimization.
Step 1: Profile the Data for Content, Structure, and Anomalies

The first step is to profile data to discover and assess your data’s content, structure, and anomalies. Profiling identifies strengths and weaknesses in data and helps you define your project plan. A key objective is to pinpoint data errors and problems, such as inconsistencies and redundancies, which can put business processes at risk.

A thorough data profiling exercise gives you a foundation for data quality success. By identifying problems up front, you avoid costly and time-consuming remediation down the road. As problems are identified, IT and business personnel investigate each data attribute and generate metadata that describes it. This metadata, or data about data, is used to cleanse data downstream or during transformation processes.

Business analysts, data stewards, and IT developers can and should collaborate on data profiling. Informatica Data Explorer helps to bridge the collaboration gap with role-based data profiling technology. Business analysts and data stewards use Informatica Analyst to assess data quality, identify anomalies, build business rules, and create scorecards.

Developers use Informatica Developer to work with the output from business users, or to generate their own data profiles. The tool gives developers greater flexibility and functionality to, for instance:

- Build, deploy, and centrally manage reusable data quality rules
- Provision data physically or virtually at any latency
- Leverage prebuilt rules for matching and address cleansing
- Reuse profiling and rules specifications across any application
- Access all data quickly to accelerate data quality projects

Figure 2 illustrates the Informatica Data Analyst interface for data profiling.

**Figure 2.** Informatica Data Analyst provides a browser-based environment for data profiling by business users.

**BANK AVOIDS $1.5 MILLION COST WITH INFORMATICA DATA EXPLORER**

Informatica Data Explorer accelerated an effort to profile data from 32 legacy applications to create a customer data warehouse at the Banco Nacional de Costa Rica. Bank officials estimate the software saved $1.5 million in labor costs that would have been necessary with manual coding.

With business and IT collaboration, the profiling exercise laid the foundation for a data quality initiative that generated accurate and trusted data from disparate sources to improve customer relationship management and profitability.

"Informatica Data Explorer is marvelous for discovering the quality of data, because the results are obtained quickly and the only limits to what you can do are in your mind—the tool always offers more," says Sergio Rodríguez, the bank’s director of databases and strategic information.
Step 2: Establish Data Quality Metrics and Define Targets

Next, you need to define metrics to measure the quality of data within your key application data fields and define individual data quality targets for each data field. The metrics should be based on the six dimensions of data quality:

1. Completeness: What data is missing or unusable?
2. Conformity: What data is stored in nonstandard formats?
3. Consistency: What data values give conflicting information?
4. Duplicates: What data records or attributes are redundant?
5. Integrity: What data is not referenced or otherwise compromised?
6. Accuracy: What data is incorrect or out of date?

You can also define custom data quality dimensions applicable to your business requirements. For example, you can establish metrics to reflect the dimension of timeliness (when data is available versus when it is expected to be available), or currency (how up to date the information is).

Tie your metrics to the business impact of data quality. For instance, correlate such business issues as stock turnover and customer shipments with data quality dimensions that can affect them (consistency and accuracy of inventory data, or duplicate customer data).

As with profiling, establishing metrics and defining data quality targets should be a collaborative and iterative effort. Informatica’s data quality solution gives business and IT a common platform to build and refine metrics that can be tracked in data quality scorecards, which may be readily shared among stakeholders by emailing a URL.

The metrics can also be viewed in Web dashboard, which offers robust drill-down and analytic reporting. Both scorecards and dashboards enable you to monitor data quality on an ongoing basis, and as you define metrics, you’ll also want to establish data quality thresholds that will trigger an email alert if breached.

Figure 3 depicts a customer data quality scorecard with metrics performance on key data quality dimensions.

Figure 3. Informatica Analyst provides a metrics scorecard that tracks performance on key dimensions of data quality.

"We wanted to invest in robust data quality tools, which were scalable and which could handle large volumes of data. We also wanted to work with one vendor who continues to support robust business solutions with business metadata management and point of entry," says Barbara Latulippe, Smith & Nephew enterprise data architect. "Time and again, it was only Informatica that could comfortably provide solutions in all these areas."

**DATA QUALITY HELPS MEDICAL SUPPLIER SAVE $1.4 MILLION IN MAILING COSTS**

The use of metrics and scorecards in Informatica Data Quality was a key ingredient in data quality success for Smith & Nephew, the London-based global healthcare company. The metrics track the company’s success in an enterprise-wide initiative to cleanse and integrate data from multiple SAP instances.

In all, the Informatica-based data quality solution helped Smith & Nephew save $1.4 million in mailing costs by cleansing customer data and reduced SKU by 50 percent by increasing visibility through metrics and scorecards.

"We wanted to invest in robust data quality tools, which were scalable and which could handle large volumes of data. We also wanted to work with one vendor who continues to support robust business solutions with business metadata management and point of entry," says Barbara Latulippe, Smith & Nephew enterprise data architect. "Time and again, it was only Informatica that could comfortably provide solutions in all these areas."
Step 3: Design and Implement Data Quality Business Rules

The next step is to define your data quality business rules—Reusable business logic that governs how data is cleansed and parsed out to populate target application fields. Business and IT teams achieve the best results when working together to design, test, refine, and implement data quality business rules using role-based functionality.

For instance, business analysts and data stewards can use Informatica Analyst to profile, analyze, and create data quality scorecards. They can drill down to specific records with poor data quality to determine impact on the business and how to fix issues. The tool enables business users to share data quality metrics and reports by simply emailing a URL to IT colleagues; it also lets business users work with developers to specify, validate, configure, implement, and test data quality rules.

IT specialists use role-based functionality in Informatica Developer to evaluate and refine data quality rules; it also includes prebuilt rules that developers can run or edit to suit objectives and capabilities for building rules from scratch.

Step 3 can also cover the development of rules for name and address matching and validation. Data stewards can kick-start projects with Informatica Data Quality Identity Match Option and its prebuilt customer data matching rules and address cleansing and validation for more than 240 countries. Developers can use Informatica Identity Resolution to search and match name and address data from more than 60 countries in batch or real time.

You can readily reuse all developed business rules across departmental applications or across the entire enterprise and its disparate data sources. Together, collaboration and reuse of business rules can slash the time and cost required for a sound and sustainable data quality project.

Figure 4 illustrates drag-and-drop configuration capabilities in Informatica Developer.

Figure 4. Informatica Developer supplies drag-and-drop capabilities to build, test, and run data quality business rules.

Collaboration is Key to Quality Success in Defense Ministry Data Migration

Collaboration between business and IT and a methodology for data profiling and quality are hallmarks of a successful migration of complex legacy data from 70 applications into a single new SAP system at the Dutch Defense Ministry.

Teams used collaboration functions in Informatica Data Explorer and Data Quality to streamline the large-scale migration and identify and correct poor data quality. Business rules incorporated into Informatica Data Quality drove cleansing, enrichment, harmonization, and integration processes.

“We elaborated the procedure for the data migration in a methodology, which describes how to carry out a quality-controlled migration in six stages,” says Jaap Timmers, the Ministry’s data migration project manager. “Business and IT now have joint responsibility for project outcomes, so they always have to be reading from the same page.”
Step 4: Build Data Quality Rules into Data Integration Processes

In Step 4, it’s time to put data quality to work as IT developers integrate the defined business rules into data quality and integration processes. How, where, and to which applications your organization applies its data quality rules will depend on the scope of your project and your data quality strategies.

Informatica Data Quality, which executes the rules, may be tactically deployed in standalone mode to function with a single business application. The solution may be configured as a preventative filter to run data quality rules “upstream,” as data enters an application.

The independent research firm Forrester Research Inc. has advocated upstream deployment to correct deficient data before it can contaminate downstream systems. “Your goal should be to use your [downstream] batch processes to supplement upstream data quality rules and processes, not as your last line of defense,” Forrester writes in a report.³

Informatica Data Quality may also be positioned to execute cleansing rules in batch mode against data once it has entered the target application, such as in a nightly cleansing process. Ideally, your data quality solutions will be deployed at the enterprise level.

Informatica Data Quality supports pervasive data quality by enabling business rules to execute as a service in a services-based architecture across any number of applications, from any point in the extended enterprise. Data quality services consist of reusable business rules that you can administer centrally, independent of applications, to perform profiling, cleansing, standardization, name and address matching, and monitoring.

Figure 5 illustrates how data quality rules, created using Informatica Data Quality, can then be deployed as broader data integration processes using Informatica PowerCenter®.

"Informatica has helped RACSA achieve what we believed was impossible,” says Manuel Pereira, RACSA project engineer. “The annual savings of $642,000 is exceptional and will help RACSA sustain its leadership in the competitive Costa Rican telecommunications market.”

**Step 5: Review Exceptions and Refine Rules**

As data quality processes are executed, most records will be cleansed or standardized and will achieve the data quality targets you’ve set. But inevitably, the quality of some data will be so poor that it will not have been cleansed, and the business rules governing data quality need to be refined.

An exception reporting process helps address and correct those weaknesses. Informatica Data Quality captures data quality exceptions and outliers and exposes them for further profiling and analysis. Role-based tools enable data stewards, business analysts, and IT developers to participate in root-cause analysis of any issues.

In addition to refining rules, you may need to edit or correct some data in place. Through its exception management feature, Informatica Data Quality allows business users to not only view but also fix data quality issues.

Exceptions may consist of duplicate records, or other contradictions or inconsistencies in data fields and attributes. Once the issue is understood, business rules may be devised or edited to address the problem at its source. By examining and addressing deficiencies, developers generate what is called the “golden record”—the proverbial single version of the truth.

Exception reporting is especially important during the initial rollout of a data quality solution and provides an opportunity to correct deficient data before it contaminates downstream applications. It’s also valuable whenever a new data source is incorporated into the data quality solution.

Figure 6 depicts an Informatica Analyst view of customer exceptions and failed records that data stewards can view, edit, and filter before they’re written to a target.

---

**DATA QUALITY IMPROVES ACCURACY FOR MARKET RESEARCH DATA PROVIDER**

Data quality is at the heart of an enterprise-level data integration effort that has enabled GfK Retail and Technology France, a provider of market research data, to improve the accuracy and reliability of information it supplies to technology and entertainment clients.

The data quality and integration effort captures point-of-sale data from thousands of merchants in 80 countries and generates more than 1,700 data files a month. With Informatica Data Quality, the volume of data errors has been reduced by 80 percent.

“Data quality is vital for GfK Retail and Technology France because it determines the quality of our customers’ decision-making,” says Fabrice Benaut, GfK director of information systems and development. “Informatica helps us control, correct, and improve data quality . . . [and] is instrumental in ensuring our customers can trust the market research analysis we provide them with.”
Step 6: Monitor Data Quality Versus Targets

Data quality should not be a one-off “set it and forget it” exercise. Continuously monitoring and managing data quality against all targets and across all business applications is essential to maintaining and improving high levels of data quality performance.

Use a scorecard or dashboard to display conformance with data quality targets and the six data quality dimensions addressed in Step 2. These monitoring tools will also reflect any custom dimensions your team has implemented and metrics that you have linked to key business performance issues.

Informatica Data Quality includes a scorecarding tool, while the Dashboard and Reports Option includes broader functionality for dynamic reporting and highly visual rendering. Customizable dashboards and reports provide high-level overviews of data quality performance and deep drill-down to assess granular issues.

Users can zoom by time to assess performance over a day, week, months, or years and be notified immediately of issues with real-time alerts. Information owners can publish monitoring results as dashboards, scorecards, and reports and share them across the enterprise.

Because data is dynamic, data quality metrics should be dynamic, as well. Look to continually adjust metrics to better reflect the impact that data quality has on key business indicators and to account for new or removed data sources.

Figure 7 shows a data quality dashboard measuring key data quality dimensions.

AVAYA OPTIMIZES BUSINESS USING INFORMATICA DATA QUALITY

Data quality is center stage at Avaya, the global enterprise network, telephony, and call center technology provider. The company has instituted a Data Quality Center of Excellence and treats data quality as a “fourth operating dimension” in the business landscape, along with people, processes, and technology. Informatica Data Explorer and Data Quality underpin the solution.

The results have been tremendous. Avaya estimates a 2,000 percent return on investment by applying data quality and monitoring across its enterprise, which includes some 400 applications and databases and 150 TB of customer, vendor, financial, and other data. A net benefit of $2 million was realized by correcting billing addresses.

“We believe that data quality is the next frontier in business optimization,” says Rich Trapp, global data quality director at Avaya. “By using Informatica within our Data Quality Center of Excellence, Avaya has access to the accurate information needed to achieve operational excellence.”

Figure 7. A data quality dashboard visually reflects conformance with six key data quality dimensions.
Conclusion

Poor-quality data doesn’t need to be a fact of life or a cost of doing business. With the right people, processes, and technology, pervasive data quality is an achievable goal for organizations in virtually any industry. A strategic and systematic methodology is an essential component that can enable your organization to optimize each resource required for data quality.

Informatica’s data quality solution provides a proven methodology and technology foundation to help your business and IT stakeholders collaborate to make data quality pervasive—ensuring that all data is complete, consistent, accurate, and current, regardless of where it resides.

With Informatica, data profiling, cleansing, address validation, matching, and monitoring capabilities are part of a comprehensive, open, unified, and economical data integration platform. The Informatica Platform lets you access, discover, cleanse, integrate, and deliver timely, trusted data to the extended enterprise—anywhere, at any time.

Build your business on relevant, timely, trusted data. Make data quality pervasive across your enterprise with Informatica’s data quality solution.

Learn More

Learn more about Informatica’s data quality solution and the entire Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the United States).

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world’s number one independent provider of data integration software. Organizations around the world gain a competitive advantage in today’s global information economy with timely, relevant and trustworthy data for their top business imperatives. Nearly 4,000 enterprises worldwide rely on Informatica to access, integrate and trust their information assets held in the traditional enterprise, off premise and in the cloud.