TDWI

MONOGRAPH SERIES

MARCH 2008

The Unique Requirements of Product Data Quality

By Philip Russom

Senior Manager, TDWI Research
The Data Warehousing Institute

SPONSORED BY

silvercreek SYSTEMS®

tdwi
THE DATA WAREHOUSING INSTITUTE

Table of Contents

Executive Summary 3
Data Quality Myths that Need Busting 4
Data quality is not one technique—it consists of many diverse techniques. 4
All data domains are not the same, so quality issues and techniques differ. 4
Data quality must address unstructured data, not just structured data. 5
Data isn't inherently right or wrong—it's suited to a business purpose or not. 6
Data quality is not a one-time action—it requires diligent repetition. 6
Data quality is seldom about perfection, because quality is relative. 6
Solution Requirements for Product Data Quality 7
Standardization 8
Verification and validation 8
Data profiling and monitoring 9
Matching 9
Data enhancement 10
Internationalization 11
Integration with other tools 11
Architectures for Product Data Quality Solutions 12
Data Quality Services 13
Recommendations 14

About the Author

PHILIP RUSSOM is the senior manager of TDWI Research at The Data Warehousing Institute (TDWI), where he oversees many of TDWI's research-oriented publications, services, awards, and events. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research, Giga Information Group, and Hurwitz Group. He's also run his own business as an independent industry analyst and BI consultant, and was contributing editor with *Intelligent Enterprise* and *DM Review* magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org.

About Our Sponsor

Silver Creek Systems puts companies in control of their product data through applications that provide data visibility and automation of key processes. Patented semantic technology and advanced workflow management deliver a full range of product data solutions from initial clean-up and standardization to enrichment, re-purposing, systems consolidation and ongoing governance across all product data domains.

The ability to systematically deliver complete, correct and consistent data from any source to core enterprise applications such as search, merchandising, PIM, MDM, ERP, PLM and GDSN substantially increases the ROI of those applications—driving revenues, reducing costs and risks.

Executive Summary

The majority of data quality software implementations have long been focused on customer data—more so than on other data domains like product data, financial data, asset data, location data, and so on. The focus on customer data helps explain why most data quality software techniques—the most common being data standardization, verification, and matching—were originally designed for customer data, whether built in-house by IT or built into a software vendor's tool. That's great for customer data, but not so good for other data domains. Customer-oriented data quality techniques and tools can be retrofitted to operate on other data domains, but with limited success. There's a need to redesign standard data quality techniques—and design new ones—that address the unique requirements of non-customer data domains.

Among these domains, the greatest need concerns **product data**, and for good reason:

Product data differs from other domains, so it has unique uses and requirements.

- **Product data represents the second-largest market after customer data.** Furthermore, product-oriented industries (e.g., manufacturing and retail) currently lag behind customer-oriented industries (financials and insurance) in the use of data quality software.
- **The quality of product data is increasingly important.** Product-oriented industries are witnessing a surge in product data volumes and product-data-driven process automation, due to the rise of radio frequency identification (RFID), global data synchronization networks, and greater software automation in general. To a lesser degree, these drivers hit all industries, since all procure goods and services, and modern procurement practices involve product data.
- **Product data differs sharply from customer data in structure and content.** Tools and techniques built for customer data's predictable syntax and patterns (expressed mostly in structured data) rarely adapt well to product data's variable syntax and nonstandard data values (commonly expressed in unstructured data).

Expecting product data to be like customer data engenders myths that need busting.

One of the greatest barriers to the cause of product data quality is perceptions. Most people's perceptions are based on experience with customer data quality or a narrow application (like name-and-address cleansing for direct mail). People need to broaden their viewpoints to encompass non-customer data domains and a greater diversity of business applications. To that end, this TDWI Monograph busts several myths associated with data quality, with a stress on product data quality's unique requirements and applications.

Standard data quality techniques don't work with product data without significant adaptation.

The most pervasive myth is that data quality is a single, monolithic entity. In reality, it consists of several diverse data quality techniques. Based on a survey from TDWI Research, the top five techniques are data standardization, verification, profiling/monitoring, matching, and integration with other tools. Techniques of lesser priority include geocoding, enhancement via third-party data, and internationalization. The common data quality techniques mentioned here are a handy mnemonic for summing up data quality priorities, and they can serve as a checklist for users in the early design phases of an implementation. Yet, each data quality technique must be adapted to a specific data domain—like product data—as this TDWI Monograph explains.

Finally, the unique requirements of product data have significant impact on the architecture of a data quality software solution. For example, product data involves a greater diversity of data sources and targets than most data domains, and many enterprise applications need to embed product data quality capabilities. Therefore, hub-and-spoke is the preferred architecture, where service-oriented spokes enable the product data quality tool at the hub to communicate with multiple sources, targets, and enterprise applications via data quality services.

Data Quality Myths that Need Busting

Despite the fast-growing adoption of data quality techniques and tools in businesses over the last 10 years, common misconceptions persist concerning software solutions for data quality. People aren't clear about what types of software techniques apply to which problems, how these work, the data domains they're designed for, and how data quality solutions can and cannot support business goals. Many myths about data quality software solutions need busting, as explained below.[1]

Data quality is not one technique—it consists of many diverse techniques.

We all use the term *data quality* as if it represents a single, monolithic approach to improving the quality of data. Even people experienced with data quality implementations talk this way, although they know that data quality is far more complicated in actual practice. Describing data quality as a single technique is misleading, because data quality is, in fact, a collection of techniques. Each technique operates on data uniquely to achieve specific data improvements or transformations.

Data quality is achieved via multiple distinct techniques like standardization, verification, and matching.

For example, common data quality techniques include various forms of cleansing, standardization, verification, and matching. To confuse matters more, some techniques have considerable overlap with others, as with the closely related techniques for deduplication, match-and-merge, and householding. And multiple techniques may be used together, the way that name-and-address cleansing invokes standardization, householding, and data enhancement techniques. Data enhancement—where external data is added to a record to complete or extend it—appends diverse types of data like geocodes, consumer demographics, and corporate firmagraphics. And, finally, data quality techniques are regularly used in tandem with techniques from related data management disciplines like data profiling, data monitoring, and data integration.

Obviously, the list of available data quality techniques is too long to repeat every time we want to talk about data quality issues and solutions. So, the single term *data quality* is convenient, and we should continue using it. Even so, let's all remember that data quality techniques are numerous and diverse. That's a strength, not a weakness, because the techniques address diverse problems and opportunities. The catch is that you have to carefully match your data quality requirements with the available solution techniques.

All data domains are not the same, so quality issues and techniques differ.

Common data domains—like customer data, product data, and financial data—each have unique requirements for data models and acceptable data values at the record and field level. This is one reason why data standards—or the lack thereof—have a big impact on the effectiveness of data quality software solutions, as well as the level of difficulty of implementing such solutions.

For example, techniques, vendor tools, and data bureau services that focus on customer data demand a recognizable customer record for the software to operate. This is seldom a problem, since best practices in modeling customer data are widely known and disseminated. That's because customer data is subject to cultural standards in how people and places are named, as well as governmentally controlled postal and telephone standards. Other customer attributes (like financials and demographics) are also defined in a universal way, though not with as much rigor as names, addresses, and phone numbers. All this ensures a predictable, base level of standardization for customer data, despite inevitable variations.

[1] For a detailed account of data quality best practices, see the TDWI Best Practices Report *Taking Data Quality to the Enterprise through Data Governance*, available online at www.tdwi.org/research/reportseries.

Customer data has a base standardization; product data doesn't.

Product data, however, doesn't have this base level of standardization. That's because just about every firm that supplies or consumes products has its own terminology, data models, and acceptable data values that it applies to describing products. Sure, there are a few standards for describing products, but these are still not used widely, and they are inherently complicated. For example, the United Nations Standard Products and Services Code (UNSPSC) currently specifies over 20,000 categories—and that's just for common commodities! Furthermore, new product data categories appear constantly. Contrast that to a well-bounded data domain—like name-and-address data—where there's a static list of approximately 330 national standards.

One of the many problems that arise from product data's lack of a base standard is the use of common data quality techniques. Most of these were designed for customer data and its predictable structures. Users and vendors have modified these to retrofit them with functions for product data, but results are mixed. Customer data has a highly predictable data syntax, and product data doesn't. So, it's best to assume that it takes special tools to deal with product data's unknown and variable data models and data values.

Data quality must address unstructured data, not just structured data.

With customer data, standardizing "avenue" to "AVE" or validating that NY is an acceptable state code are simple operations on textual information that just about any software solution can handle. Product data, however, almost always includes unpredictable strings of text, also known as unstructured data. For example, key attributes of a product are regularly described in human language, often with acronyms and abbreviations. These may be concatenated into strings, such that a tool (or human) must parse the string and extract individual data elements. (See Figure 1.)

Customer data has predictable patterns and syntax, whereas product data doesn't.

When product data records are dominated by unstructured data, the data quality software solution needs support for natural language processing (NLP) and other semantic approaches, because this level of sophistication is required for extracting data elements from unknown and unpredictable unstructured data. Note that the human-level comprehension of a semantics-driven software technique is able to locate and understand the data elements—independent of their syntax or patterns. Since data quality techniques designed for customer data rely heavily on known syntax and other patterns, their success with complex product data is generally limited.

Product data often involves strings of variable and unpredictable textual information.

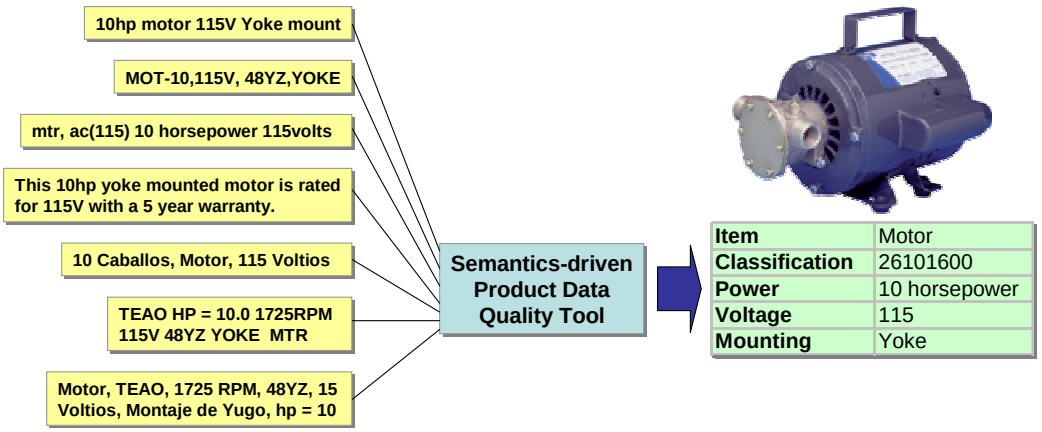

Figure 1. Source: Silver Creek Systems.

Data isn't inherently right or wrong—it's suited to a business purpose or not.

The uninitiated often jump to the conclusion that data quality is about finding and correcting defects. Indeed, some data quality techniques correct recognizable defects (like misspellings and invalid data types), but most improve data that is already relatively clean. For example, techniques like matching, householding, and deduplication find records that are not wrong, per se; they're just redundant, and so must be merged or deleted.

In the realm of product data, standardizing data (so multiple systems share the same data structures and data values) and repurposing data (to create structures and values conducive to individual applications) are common techniques applied to otherwise clean data. Product names, part numbers, and attributes of product descriptions (like ID number, weight, size) are transformed to meet the requirements of each application that uses the data. And data quality techniques may transform the same data set repeatedly as the data is needed for multiple applications. It's not that data is wrong; it's just not suited to a specific business purpose without transformation and restructuring.

Customer data attributes can be right or wrong, but product data isn't that simple.

But what does it mean to be "suited to a specific business purpose"? That depends on the business process (online sales, inventory matching, invoice matching, distribution optimization, etc.) and on the product category (color may be mandatory for shoes, but irrelevant for resistors). Rules for product attributes must be encoded in the software solution. For example, if we know that color is required—and color is present in the product record—is the color valid for the process in question? For a given application, should we describe the color as red or scarlet? Cyan or turquoise? Product data transformations depend on the context of each process or application for which it is repurposed. This is not true of most customer data attributes. For example, ZIP codes have an absolute truth—they are right or wrong. With product data there is seldom an absolute truth—there are instead business requirements to be met, and these are as varied as the data itself.

Data quality is not a one-time action—it requires diligent repetition.

Being a data steward is like being a maid: as soon as you clean up a room, someone messes it up. TDWI Research estimates that customer data degrades 10% to 12% a month, as customers move, marry, and matriculate. To stay ahead of the degradation, software must periodically monitor the quality of customer data, then operate on the data to maintain an appropriate state of quality.

Product data is equally volatile because new products appear, old ones retire, and there's a constant trickle of product data from suppliers and other sources. In addition, product data is repurposed repeatedly, which may require additional cleansing or other data quality techniques.

Data quality is seldom about perfection, because quality is relative.

It makes sense to strive for perfection in data quality when poor quality will have dire ramifications (as with financial or regulatory reporting). However, maintaining a level of quality that's relative to a specific situation is generally a more efficient use of resources and a more realistic goal, even if this demands case-by-case definitions.

For example, data about supplies that is seen and used only internally has a high tolerance for low quality, whereas product data that is seen externally by a customer (say, via an e-commerce application) needs a state of quality much closer to perfection. One of the challenges with repurposing product data is that (for some applications) data is missing, resulting in empty fields. So, an incomplete record is a common concession with product data quality.

Solution Requirements for Product Data Quality

Before we proceed, there's one more myth that needs busting:

Software solution requirements are not the same for all data quality implementations.

With data quality solutions, one size rarely fits all.

Of course, every business has unique requirements that must be met, and multiple business initiatives can have multiple sets of requirements, even when in the same enterprise. Germane to this discussion is the fact that the data domain that the solution primarily operates on—typically customer data, product data, or financial data—has a strong influence over what's required of the software solution. Since each domain has its own technical challenges—and is subject to the business requirements of disparate departments—most enterprises have multiple data quality solutions, segregated by data domain and/or department. One size rarely fits all.

Despite the diverse business and technical requirements, most data quality solutions share a handful of general requirements. To determine which data quality techniques are the top priorities for users designing a data quality solution, TDWI Research asked 750 users, "What features does your organization most need in data quality software, either homegrown or purchased commercially?" (See Figure 2.)

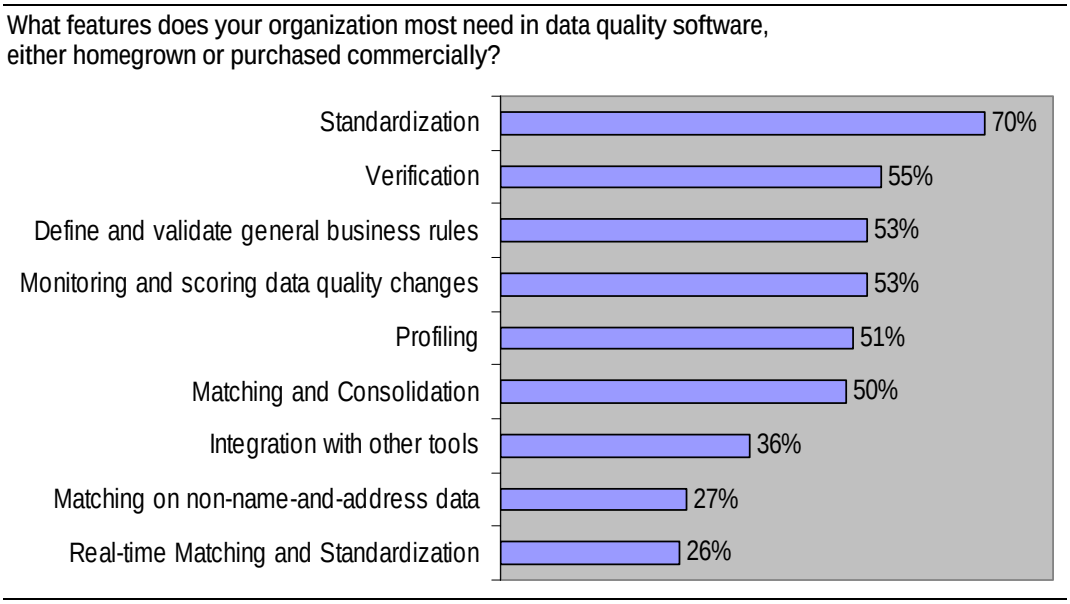

Figure 2. Based on 4074 responses from 750 respondents. Source: TDWI Research.[2]

There are multiple data quality techniques, and all must be adapted to product data.

Based on their responses, the top five techniques are (ranked from most needed to least needed) standardization, verification and validation, data monitoring and profiling, various approaches to matching, and integration with other tools. Low-priority data quality techniques (which were selected by 25% or fewer of survey respondents) include geocoding, enhancement via third-party data, and various approaches to internationalization.

[2] For a full accounting of requirements for data quality software solutions, see the TDWI Technology Market Report *Enterprise Data Quality Tools* (Q2 2006), available to TDWI Members at www.tdwi.org/research. Some of the findings of that research are restated in this TDWI Monograph, but from the viewpoint of product data quality.

The most-needed data quality techniques listed above are a handy mnemonic for summing up data quality priorities, and they can serve as a checklist for users in the early design phases of an implementation. Even so, each data quality technique must be adapted to specific data domains—especially product data—as explained below.

Standardization

Many more respondents selected standardization as a needed data quality capability than any other feature (70% in Figure 2). This isn't surprising, since standardization is fundamental to all data quality solutions, regardless of data domain, and other data quality techniques (especially various forms of matching) assume that standardization is done first. The point is to make all records in a data set conform to an established data standard, whether the standard comes from an external party or (as is usually the case) internal IT.

When product data is repurposed, there is a kind of data standard for each application and business purpose. As product data is exchanged among partners in a supply chain, it's reasonable that partners should agree to standards and comply with them. That is indeed happening, but the standards required to share the vast majority of product data are not yet defined and most are many years away. In the meantime, manufacturers must publish information to many distributors and retailers, each of whom would like the data in a form convenient for them, while distributors and retailers must cope with different format and content standards from each of their many manufacturers. And that only accounts for the data from external sources—internal sources are just as unreliable and hard to change.

If external standards cannot guide us to "perfect" data for all applications, then an alternative approach is to map disparate sources to a company standard. A product data quality solution that can perform on-the-fly standardization and prevent acceptance of poor quality data is of the utmost importance in ensuring that enterprise systems are able to operate efficiently. Conversely, lack of such a system condemns the enterprise to expensive and non-scalable manual workarounds to bring their data into compliance—a deadly cycle that is compounded by the constant changes and rapid growth of product data.

Product data cleansing and standardization are greatly facilitated by a semantic approach.

Semantics-driven data quality capabilities are crucial with product data. For a data quality tool to standardize product data, it must first be able to read and understand it. This is not trivial, since product data is often laden with unstructured data that exhibits little or no syntax or standard values. The issue is that traditional data quality tools designed for customer data—which is syntactic and pattern based—are not designed for the overwhelming complexity and variability of product data. Hence, product data benefits from semantics-driven data quality tools—especially those supporting natural language processing (NLP)—because these can extract and leverage elements in the data (especially unstructured data) that are invisible to traditional tools.

Semantics tools provide greater automation for the challenging cleansing and standardization of product data, which reduces the amount of intervention required by data stewards. The end result is a faster overall process that enables more frequent cycles in important business initiatives like new product introductions, commodity purchasing, merchandising, partnering, and so on.

Verification and validation

The terms *verification* and *validation* are used synonymously by users to mean various approaches to assuring that a data value in a field is within acceptable parameters and that a record is tagged or classified appropriately. Most verification occurs in overnight batch processing, yet most data quality vendors offer a tool or service that can verify data in real time (usually name-and-address

and financial information). Since data entry is the leading origin of data defects (and sometimes fraud, too!), real-time verification is useful for call-center and e-commerce applications.

With product data, exception management is more complex and manual than with most data domains.

Exception management plays a big role in verifying product data. Compared to customer and financial data, verifying product data has, to date, been excessively manual, due to its complex and unpredictable nature. All data quality solutions need functionality for so-called exception management (or exceptions processing), where a human user (typically a data steward) intervenes to handle records and fields that the software couldn't automatically process. With product data, exception management is quite involved, since it enables many tasks, including data value verification, product classification, data enrichment, and routing of product data records and documents. The manual exception management process for product data is slow, expensive, and error-prone, and therefore in need of better software automation.

Semantic capabilities built into a data quality tools have raised the level of exception automation considerably, and will do more in the future. But there's still a need for human intervention, especially with empty fields that only a human can fill. A data quality tool for product data needs a GUI (designed for data stewards and possibly business people like brand managers, merchandisers, and supply chain specialists) where a user can review a list of exceptions and quickly verify, approve, enrich, classify, or route them. The tool should capture rules as the steward works, so the rules can raise the level of automation. Ideally, all verification and exception management tools should have a semantic understanding comparable to that of humans, to recognize product attributes in both structured and unstructured data and process them automatically and accurately with as little human intervention as possible.

Data profiling and monitoring

Data quality users are finally taking data profiling and monitoring more seriously (51% and 53%, respectively, in Figure 2). These are two sides of the same coin, one done at design time (for data discovery and profiling), the other at run time (for monitoring data to assess its evolving level of quality). Although some users feel they can adequately perform data profiling with mostly manual methods (i.e., documenting what they discover via ad hoc queries), monitoring the quality of data over time requires the automation and consistent repeatability of a data quality tool. For the greatest usability, look for a tool that has data profiling built in, and monitors the quality of data while keeping a history illustrated with charts via a dashboard.

When the data domain is products, profiling and monitoring capabilities need to interoperate with related techniques like verification and exception management, as well as with product-specific techniques like product classification and search tagging.

Matching

Multiple forms of matching are priorities for data quality users, after standardization and verification. Its forms range from general matching (50% in Figure 2) to specific forms like customer-key management (22%) and householding (19%). Matching commonly handles name-and-address data, but also product data (27%), whether based on fuzzy algorithms or rules specific to an industry or data subject (21%). Put all these together, and matching is a large part of a mature data quality solution.

However, matching product data is quite different from matching customer data. Matching two addresses may require fuzzy logic to deal with misspellings, homophones, and so on, and there are capable solutions designed for this task, but whether they lend themselves to the different tasks and methods inherent in matching product data is uncertain.

The most common matching requirements and approaches for product data are:

- **Identity match.** The simplest form of product data matching starts with the de facto key fields of manufacturer name and part number. These are frequently misspelled or modified and may require some expertise to reconstitute them to the point where they can be used to find matching items.
- **Capability match.** Another form of matching is comparing product functional capabilities—in other words, matching items that have the same form, fit, and functional attributes. This can find identical items, and may also find reasonable substitutes or alternates. Of course, capability matching first requires extraction and standardization of the required attributes, then specialized techniques to compare required and optional (and even weighted attributes) to establish a match.
- **Relationship match.** Many products do not exist in isolation, but are assemblies, attachments, or accessories. Correctly identifying attachments and accessories is a tricky task that involves an understanding of what an item is as well as what classes of items it may be compatible with. Once an understanding is established, relationship matching can extract or infer enough information to determine what can go with what—what ink with what printer, what lens with what camera, what lamp with what lighting fixture, etc.

Accurate product classification enables successful search, procurement, spend analysis, and product recommendations.

Product classification is a form of matching that applies directly to product data. It's important that a product be matched with a product type, and these types often relate in a hierarchical classification system based around a taxonomy. The taxonomy, in turn, may structure a product catalog, data mart, or other database model. Ironically, the average raw product data record is loosely structured, yet the product it represents must be classified as a product type in a taxonomy, which is rigorously structured. To overcome this challenge, a software solution handling product data may bring to bear NLP, business rules, exception management, or a combination of these. And it should also be capable of classifying products into a number of taxonomies, both industry-standard and custom.

Matching a product with its type in the taxonomy is critical to business processes that depend on accurate comparisons of diverse-but-equivalent products:

- **Commodity procurement.** Many procurement specialists follow a fluid commodity methodology where they shop multiple suppliers with equivalent supplies, then buy according to the best price, availability, or preferred supplier status. This procurement practice isn't possible unless equivalent supplies are matched to the same classification.
- **Spend analysis.** To report or analyze your organization's spend for a specific product or a group of related products, data about the products must be classified accurately.
- **Product recommendations.** Effective recommendations for equivalent or related products in an e-commerce context demand accurate classifications in the product catalog.
- **Miscellaneous.** Product classification is a prerequisite for accurately loading product data into PIM, MDM, and BI systems. It also affects process automation, the routing of records for exception management, and the accuracy of product catalog search results.

Data enhancement

A common practice with customer data is to purchase third-party consumer demographics and append this data to customer records. This kind of data enhancement, or enrichment, is also

common with product data but occurs in a larger number of highly specialized forms. For example, an item that's missing required information can be enhanced in a number of ways:

- **Internal reference.** Information that's missing from its assigned field is often present in the record, though buried in a textual description. To be useful, this text-based data element must be identified, extracted, standardized, and loaded into the assigned field—something that can only be done effectively with a semantic-based approach.
- **External reference.** Legacy systems often contain the data elements required to complete an item record. So a product data solution should be able to extract this information from legacy systems. A common problem is that key fields do not match and so the target information cannot be found without some of the product data matching techniques described earlier.
- **Virtual reference.** The broadest and most open-ended technique, virtual references can be acquired by scraping manufacturers' Web sites or accessing an online content database (a specialized database available by subscription for certain industries). Again, finding and standardizing virtual reference data requires other standardization and matching techniques.

Most product data goes into e-commerce or procurement applications, where searchability is key.

Search tagging (or attribute standardization) is a type of data enhancement peculiar to product data. That's where metatags, parameters, and other keywords are appended to product records. These anticipate search terms shoppers and purchasers will use to assure that they find what they're looking for (and possibly related products, too), but with the shortest list of hits possible. After all, a lot of product data goes into a product catalog that's searched by mass consumers (on an e-commerce Web site) or corporate purchasers (via a procurement application). When a product data solution supports these kinds of applications, search tagging is a requirement.

Internationalization

Product data translation is made easy and accurate via NLP. Once a semantics-driven data quality tool understands a data element, it can translate the value of that element to other human languages with high speed and accuracy, especially when assisted by business rules encoded into the tool. This is important in companies that provide products to a global economy, where the product catalog needs to be updated frequently, yet also localized quickly.

Integration with other tools

Many users (36%) call a data quality tool from another tool or application:

Interoperability between a data quality solution and other systems is a requirement that affects solution architecture.

- **Data integration** naturally ferrets out data quality issues, so tools for extract, transform, and load (ETL) and other data integration techniques regularly invoke data quality tools to operate on data before loading it into a data warehouse or other database.
- **Data management applications** for MDM, PIM, SCM, and data sync all handle product data, so they may call out to a data quality tool that specializes in product data.
- **Enterprise applications** for ERP, e-commerce, call center, procurement, and so on also handle product data, though to a lesser degree, and so may need a product data quality tool. E-commerce and call-center applications regularly call a data quality server for real-time standardization and verification (26% in Figure 2).
- **Data quality tools** designed for customer data may need to call those that specialize in product data quality, and vice versa. Whether direct invocation is needed or not, data quality tools of differing specialties can complement each other in an enterprise solution.

Architectures for Product Data Quality Solutions

A number of business and technical situations affect the preferred architecture for product data quality solutions. Four of these situations stand out:

- Product data arrives from many sources, both inside and outside the enterprise.
- Product data is repurposed repeatedly, resulting in many outbound targets.
- Many applications and tools call data quality tools to embed their functions.
- Forward-looking IT departments prefer data quality functions delivered via services.

Hub-and-spoke is the preferred architecture for product data quality solutions.

Put all the above together, and the result is a hub-and-spoke architecture. (See Figure 3.) The data quality server is at the hub in the middle, surrounded by inbound spokes from data sources, outbound spokes for data targets, and bidirectional spokes for invocations from various applications and tools. The interfaces within the spokes vary from API calls and ODBC/JDBC to Web services and SOA. Product data quality services execute at the hub, but the hub may also have one or more databases that serve as a central "gold copy" of metadata, master data, and product data. Likewise, the hub may manage a repository of business rules and various other development artifacts. Finally, tools for data quality solution design and stewardship also interact with the data quality server at the hub.

Basic Technology Stack and Hub-and-Spoke Architecture for a Product Data Quality Solution

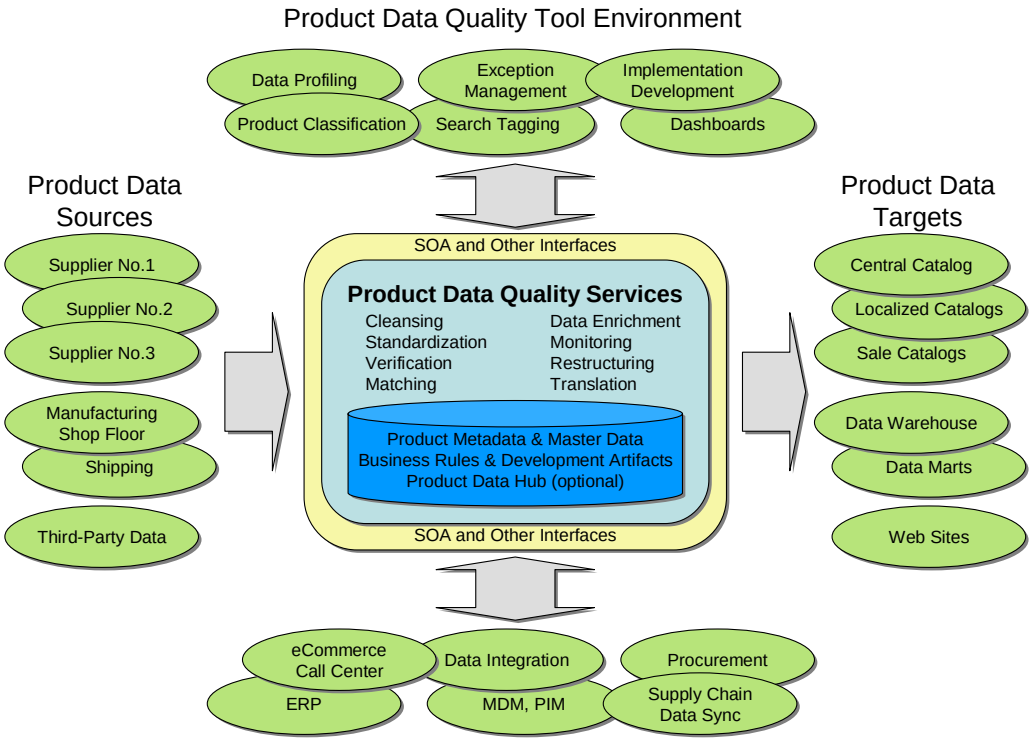

Figure 3. Source: TDWI Research.

Data Quality Services

A goal for all data quality users should be data quality services, where data quality techniques are available as services that can be called from a wide range of tools, applications, databases, and business processes. A number of benefits result from data quality services:

The point of data quality services is to improve product data everywhere it resides.

Greater interoperability. Services will help make data quality functions pervasive, but they won't replace traditional interoperability, like application programming interfaces (APIs) and standard interfaces like ODBC and JDBC. These are still required, but they should be complemented by Web services and service-oriented architecture (SOA). For Web services and SOA to be effective, a data quality tool needs an internal registry for registering data quality services and a server that can act as a hub for service requests and responses.

Modern application architectures. The loose coupling of data quality services enables a composite application methodology for data management. And an invoked service may run in real time, on demand, or in batch.

Reuse and consistency. Theoretically, any data quality technique can become an embeddable data quality service. There could also be service access to data quality design and development artifacts, like business rules, validations, standardizations, matching criteria, domain-specific terms, classifications, taxonomies, and so on. Generalizing these techniques and artifacts makes them reusable across multiple projects, systems, and business units. And their reuse fosters greater consistency in how data quality is assured and how business rules are applied.

Better data, better business. Most important, data quality services openly encourage the pervasive embedding of data quality functions into many systems and business processes, which in turn more broadly improves the data and the business.

Recommendations

To design and implement a data quality software solution that successfully addresses the unique requirements of product data, adhere to the following guidelines:

Product data differs from other domains, so it has unique uses and requirements.

Assume that product data has unique requirements. Diverse data domains have diverse syntactic structures, acceptable values, lifecycle stages, available technologies, and business applications. Embrace the diversity or risk failure in satisfying unique requirements.

Bust the myths of data quality. Most data quality myths assume a customer data mentality that misleads your approach to product data. Other myths assume that data quality techniques should be applied narrowly. Either way, bust the myths to avoid limitations and red herrings.

Don't think of product data quality as a matter of right and wrong. Data that is high quality for one application may be useless for another. Be sure your approach to product data quality allows for fast and easy repurposing of product data into any required format or standard.

Product data's reliance on unstructured data demands natural language processing.

Recognize that product data is fundamentally different from customer data. Customer data has predictable patterns and syntax (expressed mostly in structured data), unlike product data's variable syntax and textual data elements (commonly expressed in unstructured data). Tools and techniques designed for customer data need serious adaptation before they apply to product data.

Look for tools that support natural language processing (NLP). Semantics-driven software is able to extract the complex data elements buried within product data, independent of their syntax.

Automate exception management as best as you can. Due to its complexity and variability, product data quality requires large doses of exception management. To reduce this, look for semantics-driven tools that capture business rules heuristically as a steward manages exceptions.

Techniques key to product data quality include automation for exception management, product classification, search tagging, and fast repurposing.

Monitor the quality of product data to assure continuous improvement. As with all data domains, product data benefits from quality metrics displayed in an intuitive dashboard.

Ensure you understand your matching and enrichment requirements. These are common requirements for all data, but have very different implications when dealing with product data.

Classify products carefully. This is a prerequisite for business practices like commodity procurement, spend analysis, product recommendations, and effective product catalogs.

Enhance online product data stores with search tagging and attribute standardization. This helps online shoppers and corporate purchasers find exactly what they need, plus related products.

Organize product data so it's easily repurposed. A beneficial byproduct of NLP is that—once data elements are understood semantically—they are easily restructured, translated, and localized.

Product data quality solutions need a hub that supports many interfaces and SOA.

Deploy a hub-and-spoke architecture for product data quality. A hub managed by a product data quality tool is required because product data comes from many sources, it's repurposed to many targets, and many enterprise applications, tools, and databases need to call the tool.

Adopt service-oriented architecture (SOA). Although traditional interfaces aren't going away, SOA enables loosely coupled data quality services to participate in composite applications.

Gain the widest reach and deepest impact through product data quality services. These help make data quality functions pervasive in a wide range of applications, which in turn broadly improves product data and the business processes that depend on product data.