



Published in DM Review in October 2006.
Printed from DMReview.com

Product Information Management - Forcing the Second Wave of Data Quality

by Martin Boyd

Summary: *This article appears in the DM Review Special Section focusing on MDM, CDI and product information. Product data presents challenges when it comes to quality issues. Does it need to be handled differently than customer data?*

Data quality tools have been around for some time, but the requirements of master data management (MDM) and product information management (PIM) are forcing a new wave of innovation and technology breakthroughs. The need to share information across the enterprise and supply chains is driving data from legacy application silos to be increasingly exposed and shared. This reveals massive inconsistencies and incompatibilities - hence, the recent interest in MDM and related technologies that promise to unify and synchronize this disparate data and deliver a single version of the truth.

Before a single version of the truth can be maintained, it must first be created, and that is more easily said than done. Loading an MDM system with inconsistent, as-is data risks a rude lesson in "garbage in, garbage out" and possibly the failure of the entire initiative. Disparate data must first be understood, then cleansed, enriched, standardized and restructured. Without this, records can't be reliably loaded, compared and matched across the enterprise.

For customer data integration (CDI), this task is relatively straightforward. CDI is the customer-focused subset of MDM and has benefited from the first wave of data quality technology, which has proven to be very effective at solving name and address problems. These traditional tools use algorithms and heuristics to correct keyboard entry errors, phonetic misspellings, alternate name forms (Robert and Bob), invalid ZIP codes, household variations and similar challenges. Such technologies are syntax and pattern based. They are quite mature but still remain primarily focused on name and address issues, despite repeated attempts to broaden their scope.

For PIM, the challenges are much tougher. This product-focused branch of MDM has demonstrated that syntax-based data quality tools perform poorly when faced with the overwhelming complexity and variability of most product data. Match rates seldom exceed 50 percent. So, does product data need to be handled differently for successful PIM? To answer that, let's first look at what makes product data so challenging.

What's Different about Product Information?

Name and address data is typically riddled with errors that need to be detected and corrected. Product information, on the other hand, is rich with meaning that must be comprehended. The handling rules are different for product information.

The truth can exist in multiple, equally valid forms. There is no single valid way to describe products and virtually no standards to guide us. For example, "10hp ac motor" and "MTR, alternating, 10 horsepower" may both be equally accurate descriptions and perfectly understandable to the human eye, but neither description will be very usable by a PIM system that requires structured and attribute-based information, such as "item = Motor; power = 10 horsepower; type = alternating current."

Product-specific attributes are essential. Most product descriptions contain many valuable attributes, but they are usually embedded in random and unpredictable ways and are difficult to extract so they can be used effectively. This problem is compounded by the fact that every product category has a different set of attributes - for example, shoes have different attributes (style, size, color) than motors (power, type, frame size), and motors have different attributes than digital cameras (resolution, zoom range, memory type). The only way to reliably understand and describe a product is to isolate and extract its attributes, a task which traditional tools that rely on identifying syntactic patterns do poorly because of the complex, variable and ambiguous nature of product data.

There are many, many domains to master. For data quality purposes, name and address data can be considered as a single domain - within a single country, the rules do not change. Product data is far more diverse. For example, the UNSPSC (United Nations Standard Product and Services Code) schema lists more than 20,000 categories, and for many, even this is not sufficiently detailed. Product information has a huge diversity of subject areas, each with its own set of attributes, terms, grammar and abbreviations. Every one of these domains must be absorbed and understood if the information is to perform reliably in the system.

There are many different output formats. The job is not done even when your product information has been properly understood. Different descriptions may be required for the Web site versus the inventory system or even for Customer A versus Customer B. Different attributes and classifications may be needed for search versus business intelligence reporting. Additionally, there's the daunting task of having to translate product information into many different languages for your global audience.

Philip Howard, the director of technology research for Bloor Research, has investigated different approaches to data quality. In a recent research article, he notes, "Product matching is an order of magnitude more complex than conventional name and/or address matching. While there are some relatively simple product-based environments in which traditional methods can work well, in more complex situations, success rates are seldom above 50 percent."¹

Of course, there has always been one solution with a much higher success rate in dealing with complex and unpredictable data: manual review and editing. Unfortunately, this solution is also slow, expensive and not very scalable.

The Semantic Shift

It is time for a paradigm shift. Human beings can easily understand and transform the complex information required by a PIM system. Traditional, syntactic, pattern-based tools choke on the complexity and variability of that same data. Are humans processing the data in some fundamentally different way? The answer is a resounding yes - the human brain is known to focus much more heavily on the semantics (meaning) of data, as opposed to the syntax (pattern) of the data. That distinct approach makes a world of difference when it comes to deciphering complex and variable data such as product information.

A growing number of Fortune 1000 companies are having great success using semantic technologies to process their product information. They're finding that semantically savvy solutions are quick to deploy, highly flexible and reusable, and can be operated and maintained by business users with little or no IT support. This success is the early evidence of a second wave in data quality innovation.

Catching the Wave

Until recently, customer data has been the squeaky wheel in the data quality space. Now, the tougher challenges posed by product data are forcing change and driving innovation. Look across the enterprise. The same kinds of challenges exist anywhere complex data must be understood and matched: in inventory consolidation, Web and catalog publishing, spend classification, product-line reporting or supply chain integration. The potential use cases stretch to the horizon.

The second wave of data quality solutions will reflect a thorough understanding of data semantics to greatly enhance and extend the existing data syntax tools. Second-wave solutions will be very flexible and tolerant of data complexity, making them highly adaptive and able to rapidly assimilate new data domains.

The elusive promise of a single version of the truth is about to be realized. Complex product data is forcing a fundamental rethinking of what we mean by data quality, and in the process, it is removing the last few hurdles to successful PIM and MDM deployments.

Reference:

1. Philip Howard. "Silver Creek Systems - not statistically challenged." IT Analysis/Bloor Research, 24 March 2006.

Martin Boyd is vice president of marketing at Silver Creek Systems, a product data quality company. He may be reached at mboyd@silvercreeksystems.com.

Copyright 2007, SourceMedia and DM Review.