



www.silvercreeksystems.com

Automating Product Data Quality for Trouble-Free PIM and MDM

*Product Data Quality & Governance Can Create
PIM-Perfect Data and Maximize PIM-MDM Performance*

SILVER CREEK SYSTEMS: WHITE PAPER

Silver Creek Systems - 10385 Westmoor Drive - Suite 225 - Westminster, CO 80021 - 720.304.9828

December 2006

EXECUTIVE SUMMARY

Some companies deploy Product Information Management (PIM) or Master Data Management (MDM) systems and are dismayed to discover that their underlying data is far more inconsistent and unusable than they suspected. Gartner Research put it succinctly: "Most companies are in a state of denial about their data quality issues..." These unaddressed quality issues are usually decades old and can be significant roadblocks to achieving the many important benefits that justified the PIM-MDM initiative in the first place... So, what can be done?

The answer lies in understanding the very different technical challenges that fall under the label of Data Quality. Historically, Data Quality has primarily consisted of "People," "Product" and "Financial" data. To a large degree, financial data was successfully standardized by ERP systems and people data (customers, partners) by CRM and CDI (Customer Data Integration) systems. Now, it's time to tackle and solve the remaining quality issues primarily associated with product data.

"Product Data Quality" (PDQ) is the term that has emerged as the solution for this most recent roadblock to enterprise integration. PDQ is the key to achieving the elusive corporate goal known as the "single source of truth" that underlies both PIM and MDM as well as most aspects of the Information Supply Chain. Seeing product data as a problem area isn't all that surprising when you consider the following:

- On average, PIM systems must combine 5-6 sources of product data and synchronize up to 25 disparate systems that rely on product data
- More than 90% of the content on e-commerce websites is product information
- Product information permeates the entire supply chain, from inventory to spend management

Of course, before a PIM-MDM system can become a reliable single source of truth, it must first be loaded with high quality data. This is definitely easier said than done. Loading it with inconsistent, as-is data risks a rude lesson in "Garbage-In, Garbage-Out" and can even delay or threaten the entire initiative... hence the quest for 'PIM-perfect' data that will be accurate, complete, attribute-rich, standardized, classified and translated. Historically, it's been a daunting task, but now semantically-savvy technologies are removing the last few barriers in the process.

ROOT CAUSES

When companies attempt to merge product data from such disparate sources as customer requests, supplier updates, or ERP, PLM or ECM systems, the resulting problems are easy to see: 'master' product data that is fragmented, partial and often redundant.

Problem data like this makes it impossible to consolidate spending across suppliers, get accurate inventory reports, publish standardized information, or get reliable answers out of business intelligence. This, in turn, can delay time-to-market and may even expose the company to unacceptable compliance risks.

What's needed is an automated solution that will take different versions of a single product record and blend them into a Single Source of Truth (SST) record for the enterprise. This SST record is not necessarily in a format that will ever be published in its entirety, but this record encapsulates the company's essential understanding of what an item is and how it can be described in new ways without losing its underlying identity.

Companies have traditionally used a brute force approach to achieve master product data – a mix of extractions, pattern-matching routines, scripting, data merges and a strong dose of manual effort, all in an effort to get data loaded into the master repository. The results have generally been mediocre.

WHY ARE THINGS SO OUT-OF-SYNC?

Every application uses its own unique set of rules and logic to handle data, and that's why even two seemingly similar systems are never in sync. Universal standards for data handling might change that in the distant future, but until then we will all be forced to reconcile data that is inconsistent, missing, or buried in databases, documents, email, or on paper.

Since applications won't generate synchronized data any time soon, what's needed is a dynamic and real-time process to synchronize that data as it moves between systems.

TRADITIONAL SOLUTIONS DON'T WORK WITH PRODUCT DATA

Product data is far more challenging than "People" or "Financial" data, and poses some unique problems. That's why traditional, syntax-based data quality tools perform poorly when faced with the overwhelming complexity and variability of most product data.

"Product matching is an order of magnitude more complex than conventional name and/or address matching," says Philip Howard, the director of technology research for Bloor Research. "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] Why is this so?

Take customer data integration (CDI), for example, a classic People-related data quality challenge. CDI 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 (ex: Robert and Bob), invalid ZIP codes, household variations and similar challenges. They are quite mature but are still mostly limited to name and address issues, despite repeated attempts to broaden their scope. That's because these tools are syntax- and pattern-based and cannot deal with the inherent complexity, variability – and sheer unpredictability – of most product data.

Of course, there is one method that performs quite well with complex product data — manual review and editing. Unfortunately, this is also a method that is too slow, expensive and error-prone to be sustainable, yet it gives us some clues on how to tackle the product data problem.

Human beings can easily understand and transform the complex information required by PIM systems but 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 "yes" — the human brain is known to focus more heavily on the semantics (meaning) of data, as opposed to the syntax (pattern) of data. The semantic approach makes a world of difference when it comes to deciphering complex and variable product information.

From an IT perspective, it's clear that semantically-based Product Data Quality (PDQ) solutions are able to extract and leverage elements in the data that are 'invisible' to traditional tools. The DataLens System™ is a leading example of semantic-based solution and later in this paper, we'll discuss where and how it can be applied. First, however, we need to better understand what makes product data so challenging compared to People or Financial information. Let's examine the differences...

WHY IS PRODUCT DATA SO DIFFERENT?

In the CDI example above, traditional pattern-matching tools are designed to "detect and correct" errors in name & address information. Product information, on the other hand, is rich with meaning that must be comprehended. That difference changes the rules for handling 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."

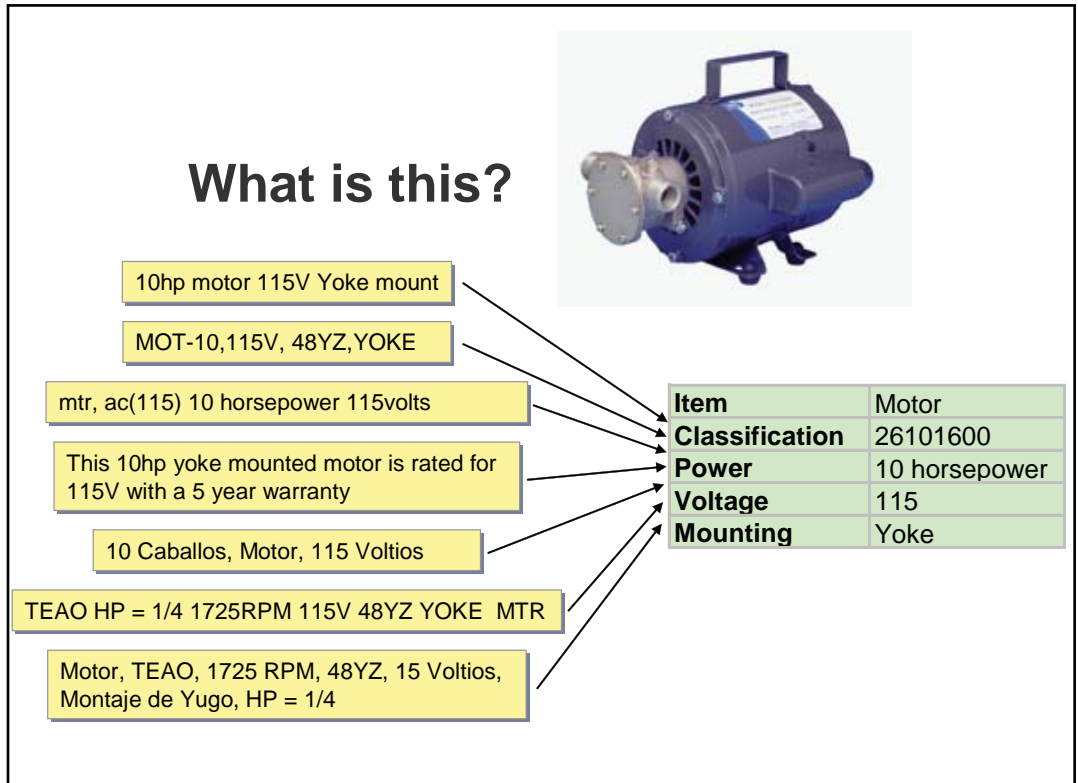
DIFFERENT RULES FOR PRODUCT DATA

Truth can exist in multiple, equally valid forms

Product attributes are essential and must be extracted

There's an ever-expanding universe of product domains

Deciphering all this requires the semantic processing model of the human brain



Product information can exist in many equally valid forms. The information may be attribute-rich, but is often cryptic, unstructured and unpredictable

Product-specific attributes are essential. Most product descriptions contain many valuable attributes, but these are usually embedded in random and unpredictable ways and are difficult to extract. 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 pattern-based tools do poorly because of the complex, variable and ambiguous nature of product data.

There are 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 28,000 categories, and for many companies, 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 in context if the information is to perform reliably in the system.

Many different output formats are required. The job is not done even when your product information has been properly understood. Different descriptions may be required for the website 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 an entire global audience who may need your information in many different languages.

WHAT IS PRODUCT DATA QUALITY?

Simply put, Product Data Quality has the goal of creating and maintaining product information in a well-defined, standardized and extremely flexible state so that it is fully 'fit for use' anywhere in the enterprise, on-demand.

Product Data Quality needs to be an enterprise-wide function, not just a technology for data cleansing and scrubbing. It's a much more involved process that focuses an organization's resources on addressing quality issues at the source vs. after the fact, in all of the organization's data warehouses and analytic/reporting platforms.

That said, a powerful and effective technology for assuring Product Data Quality is obviously the first step on the road to overall governance. Automated and semantically-based PDQ solutions are proving themselves to be a valuable addition to the IT arsenal. They do much more than just "detect and correct" — they offer human-level comprehension and flexibility in discovering and leveraging the meaning of product data, which makes that data easy to move, match and transform:

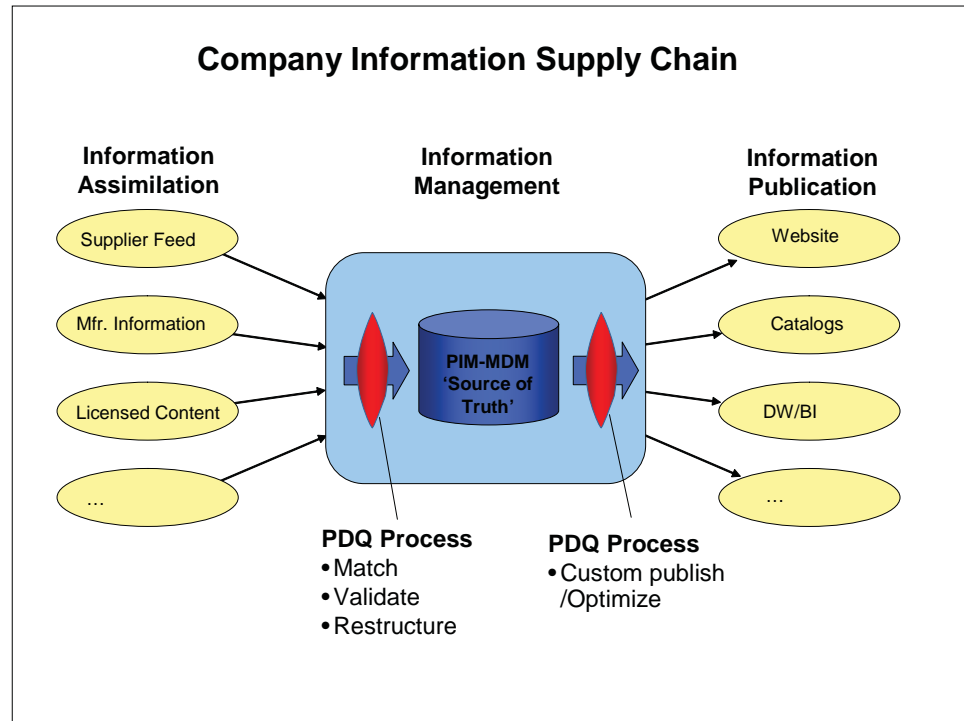
"The whole point of semantic integration is to cleanse and integrate data by understanding the meaning of the elements, not just their syntax or patterns. An added benefit is that once the meaning of something is understood, it can be translated into other languages without 'semantic loss.'"^[2]

PDQ IN THE INFORMATION SUPPLY CHAIN

Any process that relies heavily on product-related information is a candidate for a Product Data Quality process, but it's especially critical in certain fields, including Retail, Distribution, Manufacturing and online Search & Navigation, all of which rely heavily on a product information supply chain.

For such companies, the Information Supply Chain conveys vital information about the goods being offered for sale, and usually this is critical to the items being found and bought. Any blockages or breakages in the chain will have a serious negative

effect on revenues. Conversely, a streamlined Information Supply Chain can be a source of competitive advantage, customer satisfaction and increased revenues. To avoid blockages and ensure that the right information is available in the right form at the right time, Product Data Quality and Governance should be deeply integrated within the Information Supply Chain.



The Information Supply Chain of a single company

- **Information Assimilation** – In the typical product information supply chain, product information is assimilated from many different sources – from manufacturer websites to licensed content feeds. Often this information is poorly structured, but even when structure exists, it is unlikely to meet the needs of the internal systems and the PIM-MDM systems. Before being accepted into the organization, the incoming information must be understood, matched, validated and restructured – a Product Data Quality Process.
- **Information Management** – Once accepted, product information must be synchronized across many systems. This is a key reason to have a PIM-MDM system in the first place, and that system is the synchronization tool of choice, except when unusually complex restructuring or matching is required as part of the process.
- **Information Publication** – A key step in the Information Supply Chain is publishing the information out of the PIM-MDM and related systems, for consumption by other systems. One of the most important consumers of this information is the e-commerce website which must have product information optimized for

both search and navigation technologies as well as for human queries. Manufacturers and distributors may also need to publish the information in a large number of custom forms as required by downstream retailers who sell their goods.

Of course, similar needs apply in many other industries. The requirement to assimilate, manage and publish disparate product information between businesses and systems is extremely widespread.

WHAT IS PRODUCT DATA GOVERNANCE?

Governance is as much about business strategy and process as it is about technology – assuring a fundamental change in business processes so data stays cleansed and standardized and is not just a one-time event.

On the simplest level, Product Data Governance is the enterprise-wide enforcement of content standards, to achieve and maintain a reliable single source of truth within the enterprise. This is a broad and important topic covering methodology, process, and technology areas that cannot be completely tackled by any single solution. On the other hand, having a technology that can enforce compliance to content standards is an enormous step in the right direction and can be the foundation for other methodology and process initiatives.

The technology component of product data governance should include the real-time ability to validate manual entries and batch loads as well as create data scorecard metrics for quantifying the completeness and validity of each data source.

This commitment, coupled with an effective, real-time process that keeps product data synchronized between systems, allows the organization to repurpose data on-demand, to match its ever-changing needs. Again, semantically-based solutions can play an important role in achieving the 'PIM-perfect data' that facilitates overall governance.

WHAT IS "PIM-PERFECT" DATA?

PIM and MDM systems perform best when the data loaded into them is an exact match to the system's internal data standards (schema). Although some existing enterprise data may be labeled as "standardized," if it's not an exact match to your system's internal standard, it won't be that useful in developing a 'single source of truth.'

In any case, most product data is far too inconsistent and unstructured for reliable use in a PIM-MDM system – loading it as-is will undermine the value of the system and produce partial or flawed results... the very thing you wanted to avoid.

To qualify as PIM-perfect, the process must be capable of producing product data that is...

- *Standardized* - Text descriptions and product attributes must use consistent terminology, abbreviations, word order, punctuation and units of measure. In most as-is product data, key attributes are often buried in descriptive text – these must be extracted and standardized. This is the difficult but necessary task that allows product records to be matched by comparing functional attributes.
- *Localized* – Global companies may offer product information in many languages, but only have it available in one language. Conversely, they may have it in only one language but need to publish it into many languages, for their expanding global audience. This may be done when the data is imported to the PIM system, or when it is published.
- *Enriched* – Data enrichment can take many forms – extraction and population of standard attributes, classification of the item into one or many taxonomies, or by augmentation of missing information by referencing third party sources.
- *A Perfect Match for the PIM System* – Even where ‘standardized’ data exists, is it the right standard that will allow easy loading, trouble-free operation and peak performance for the PIM system? Only an exact match of the incoming data to PIM system standards can guarantee this.

The DataLens System meets all these requirements and can take product data from any source and deliver it in a PIM-perfect format, thus solving one of the thorniest problems in PIM-MDM deployments. In fact, if data standards (schemas, validation rules, etc.) have already been loaded into the PIM or MDM system, the same standards can also be loaded into the DataLens System, to assure “PIM-perfect” data from the outset.

THE SEMANTIC SHIFT

The externalization of enterprise data has put a spotlight on the problems inherent in product data – immense variability, conceptual complexity and a relative lack of standards. Clearly, trying to predict all the permutations of product data is a truly impossible task, which rules out pattern-based tools as a viable option. A completely new approach is needed.

Automated solutions for Product Data Quality are emerging and helping a growing number of Fortune 1000 companies to have increased success in leveraging their product information. These companies are 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 heralds a second wave of innovation in the data quality field.

The DataLens™ System, from Silver Creek Systems, is a noteworthy example of this recent wave of innovation. Let's examine this semantically-based solution for Product Data Quality and Governance.

THE DATALENS SYSTEM

The DataLens System is software that blends semantic modeling, expert systems, artificial intelligence and learning theory. It captures human-level rules for data comprehension, then applies these rules to cleanse enterprise data streams.

Every DataLens System ships with a library of pre-built domain- and industry-specific rules ("data lenses") that assure a rapid deployment. Data lenses are easily shared and reused across the enterprise and can be linked together to solve complex data problems in multiple domains.

Should the need arise, custom data lenses are easily built – the system uses a "watch & learn" process to observe as human subject matter experts (SME) interact with the problem data, then it infers the underlying rules of semantic understanding and creates a new lens that transforms the data as a human would... only much faster, more consistently and on-demand.

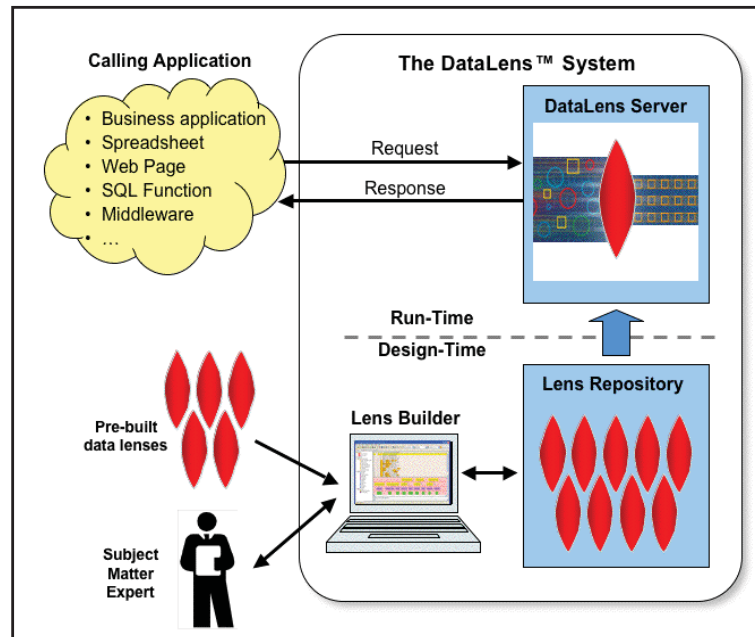
The system consists of three components:

DataLens Builder - the design module that creates or customizes the 'lenses' that will transform the data. Operators don't need any special technical skills, just expertise in the business domain of the data being processed.

DataLens Server - the run-time module that operates as a scalable, standards-compliant web service, delivering cleansed product data on-demand.

Data Lenses - Data Lenses are files that store the captured expertise and semantic rules for deciphering and transforming product data in one or more domain or industry. Powerful pre-built lenses that handle most challenges come with every system, and it's easy to customize these lenses or build new, specialized lenses from scratch.

The DataLens Builder is quite different from any traditional approach to product data quality. This is where a human expert – typically a business user – interacts with small samples of the data to 'teach' the system how to understand the data semantically and shows the system what that data should look like when restructured for other uses. Its simple drag-and-drop interface quickly defines the contextual relationships within the data without the need for coding or IT assistance.



Once the context has been established, the entire dataset can be restructured, as required. This step includes:

Standardization – of descriptions and attributes to any content standard
 Classification – to any taxonomy

Translation – from any language, to any language

Match & de-duplicate – use workflows to compare standardized versions of data

The **DataLens Builder** detects exceptions and flags them for review and additional rule-building. Very quickly, the system achieves high levels of data comprehension and accuracy, with no suspect data being passed downstream to users. Equally important is the system's ability to extrapolate solutions to new domains and semantic challenges, based on its prior learning experiences.

The DataLens Server is the run-time component of the system that performs on-the-fly transformations of "as-is" product data into the "as-needed" formats required across the enterprise.

As a Web-Service, this server is built on open standards, making it simple to integrate. Because it has no persistent storage, it also scales easily with simple hardware upgrades. It runs very efficiently on its host platform, either on-the-fly or in batch mode (synchronous or asynchronous) and blends seamlessly with other background processes.

In addition, the DataLens System uses a series of 'maps' to manage its processes in a fully integrated, closed-loop environment that can be used to reflect complex workflows for data governance and integration.

REAL-WORLD RESULTS

Silver Creek Systems has customers in distribution, manufacturing, e-commerce and other product-focused industries. The company's DataLens System typically deploys in days-to-weeks, with significant improvements in overall Product Data Quality being seen in 60-90 days. As a result, the ROI for this solution is unusually rapid.

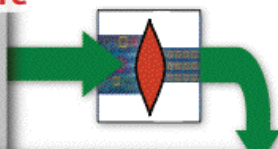
The following illustration shows a typical 'Before' and 'After' scenario for the DataLens System:

**Real-World Examples:
Electronic Components / Resistors**

SUPPLIER'S DESCRIPTION

RESPCF AX300OHM1/4W5%
220 ohm CF AX 0.5W 5% res
array 16 pin 85 ohm 5% re
NTWK 330 OHM 8 P 2%
CF AX 75 OHM 1/4 W 5% RESP
NTWK 25P 220 OHM RES
RES MF AX 226 OHM 1/4 W 1%
RES TF CH 100 OHM 1/10 W
RESS2.21 OHM 1% 1/10 W TF CH
RESS 75 OHM 1% TF CH 1/10
RES 20 OHM 1 W 5% CARBON
RES array 16 PIN 10 OHM 5%
ARRAY 100 ohm 16 pin SMD
ARRAY SMD 16 PIN 62
cf ax1000ohm5W2%
ARY 51 OHM 8 PIN
RES 226OHM MF AX
TF CH 100 OHM 1/1
RES2.21 OHM 1% 0
75 OHM 1% TF CH

Before



After

Resistance	Power	Tolerance	UNSPSC	FSC	English Description
300 Ohm	0.25 Watt	5%	32121609	5905	Resistor 300 Ohm 5% 0.25 Watt Axial Carbon Film
220 Ohm	0.5 Watt	5%	32121609	5905	Resistor 220 Ohm 5% 0.5 Watt Axial Carbon Film
85 Ohm		5%	32121607	5905	Resistor 85 Ohm 5% 16 Pin Array
330 Ohm		2%	32121607	5905	Resistor 330 Ohm 2% 8 Pin Network
75 Ohm	0.25 Watt	5%	32121609	5905	Resistor 75 Ohm 5% 0.25 Watt Axial Carbon Film
220 Ohm			32121607	5905	Resistor 220 Ohm 25 Pin Network
226 Ohm	0.25 Watt	1%	32121609	5905	Resistor 226 Ohm 1% 0.25 Watt Axial Metal Film
100 Ohm	0.10 Watt		32121609	5905	Resistor 100 Ohm 0.10 Watt Thin Film Chip
2.21 Ohm	0.10 Watt	1%	32121609	5905	Resistor 2.21 Ohm 1% 0.10 Watt Thin Film Chip
75 Ohm	0.10 Watt	1%	32121609	5905	Resistor 75 Ohm 1% 0.10 Watt Thin Film Chip
20 Ohm	1.0 Watt	5%	32121609	5905	Resistor 20 Ohm 5% 1.0 Watt Carbon
10 Ohm		5%	32121607	5905	Resistor 10 Ohm 5% 16 Pin Array
100 Ohm			32121607	5905	Resistor 100 Ohm 16 Pin Array Surface Mount Device
62 Ohm		5%	32121607	5905	Resistor 62 Ohm 5% 16 Pin Array Surface Mount Devic
1000 Ohm	5.0 Watt	2%	32121609	5905	Resistor 1000 Ohm 2% 5.0 Watt Axial Carbon Film
51 Ohm		5%	32121607	5905	Resistor 51 Ohm 5% 8 Pin Array
226 Ohm	0.25 Watt	1%	32121609	5905	Resistor 226 Ohm 1% 0.25 Watt Axial Metal Film Surfa
100 Ohm	0.10 Watt		32121609	5905	Resistor 100 Ohm 0.10 Watt Thin Film Chip
2.21 Ohm	0.10 Watt	1%	32121609	5905	Resistor 2.21 Ohm 1% 0.10 Watt Thin Film Chip
75 Ohm	0.10 Watt	1%	32121609	5905	Resistor 75 Ohm 1% 0.10 Watt Thin Film Chip

Additional information on Silver Creek Systems and the DataLens System can be found at <http://www.silvercreeksystems.com>.

CONCLUSIONS

The PIM and MDM markets are growing rapidly, fueled by the need for consistent and synchronized product information, both within and between companies. But like most IT initiatives, PIM and MDM must avoid the 'Garbage In, Garbage Out' trap, because implementers are finding that it requires a lot of time and effort to massage their data in order to make it useful as a 'single source of truth.' This can – and often does – delay deployments and seriously undermine project ROI.

Clearly, an automated Product Data Quality solution is required to take messy and inconsistent data and put it in a PIM-perfect form so it can be easily loaded and shared. Unfortunately, traditional pattern-based tools quickly choke on the complexity and variability of most product data, so a new approach is required.

Silver Creek Systems has developed a solution that can automate Product Data Quality processes using semantic-based technology that easily handles unpredictable product data streams and transforms them into PIM-perfect data in real time. This not only reduces system deployment times and ongoing maintenance costs, but can be the difference between project success and failure.

As the pressure for supply chain integration intensifies, both PIM and semantic-based Product Data Quality solutions will become an accepted part of the enterprise landscape.

As one analyst put it, "the transition from traditional pattern-based tools to a semantic-based tool has been eye-opening for business and technology users alike, and represents a significant competitive advantage..."^[3]

References:

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

[2] Jim Murphy, Rob Bois and Eric Newman, "Taking the PIM Path on the MDM Journey" in a July 2005 report by AMR

[3] Neil Raden, "Semantic Integration: Tapping the Full Potential of Enterprise Data," January 2006.



www.silvercreeksystems.com

Automating Product Data Quality for Trouble-Free PIM and MDM

Silver Creek Systems
10385 Westmoor Drive, Suite 225
Westminster, CO 80021
U.S.A.

Worldwide Inquiries:
Phone: +1 720-304-9828
Fax: +1 720-304-0531
www.silvercreeksystems.com

Silver Creek Systems, DataLens, and Product Data Quality are trademarks of Silver Creek Systems, Inc.
Copyright © 2006 by Silver Creek Systems; all rights reserved.