Enterprise Data Record Level Semantic Model of Enterprise Content

By Ed Green, Ph. D.
Chief Technology Officer

Silver Creek Systems, Inc.
www.silvercreeksystems.com

ABOUT THE AUTHOR

Ed Green leads a multi-disciplinary team at Silver Creek to develop new data integration techniques that draw on everything from semantic modeling to expert systems and artificial intelligence. His 30 years of experience includes service as the VP of development and COO of Cadis Inc., a parametric search engine company, and GrafTek, a solids modeling CAD/CAM company.

Ed can be contacted at egreen@silvercreeksystems.com

ABOUT SILVER CREEK SYSTEMS

Silver Creek Systems offers a breakthrough technology solution that takes inconsistent data from disparate sources and restructures it on-the-fly to make that data truly usable – standardized, localized and enriched – all across the enterprise.

Traditional tools perform poorly with complex and variable data – especially product data. The Silver Creek breakthrough is in using semantic identification to extract & leverage data elements that are 'invisible' to these traditional tools.

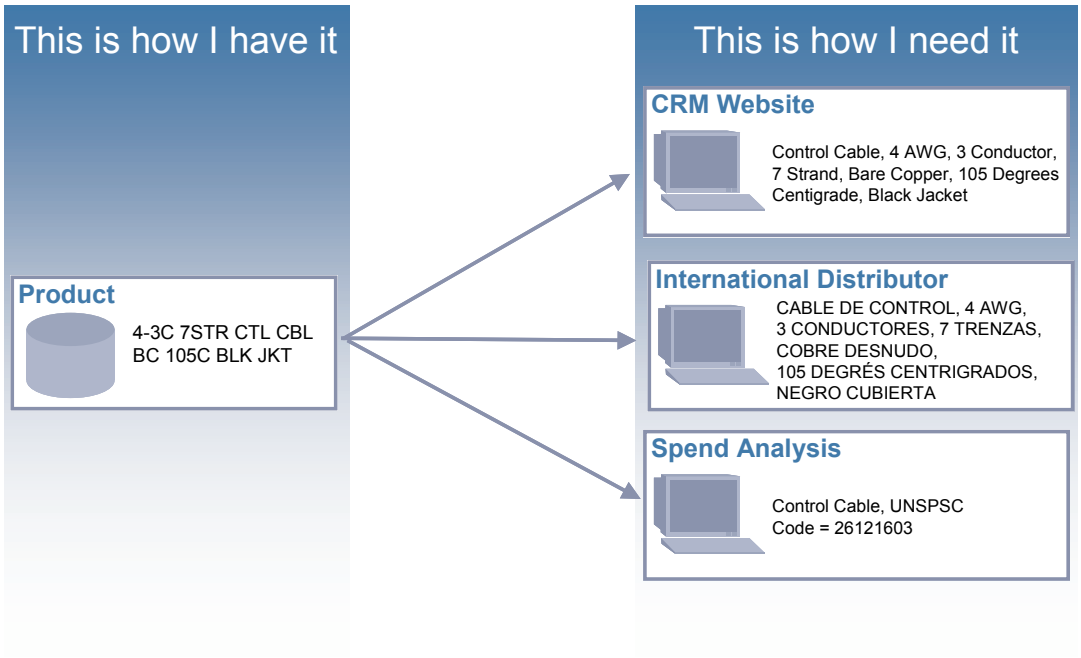
The Silver Creek solution makes even the most complex data easy to restructure, reuse and repurpose, for rapid use whenever and wherever it's needed.

For more information visit www.silvercreeksystems.com/semantic or email inforequest@silvercreeksystems.com or call 720 304 9828

The Enterprise Data Challenge

Data quality initiatives, business intelligence, data warehouses and data marts all attempt to make enterprise data – data that is the core of a firm's products and services – more valuable and more timely. Systems that create "common views" of customers, suppliers, partners, products and processes will continue this theme of exploiting the increasing value of enterprise data. In spite of all of these various initiatives and forays, the root problem usually comes down to the following data conundrum:

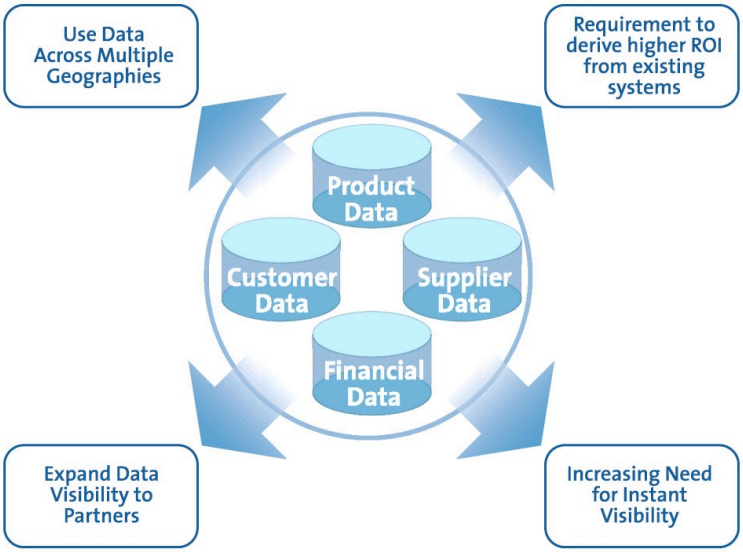
The data exists in various systems one way but it is needed in other ways. In the figure below, highly abbreviated product data needs to be expressed in multiple ways. The problem then becomes how to do this in a manner that is compatible with the exploding uses, quantities and increasing criticality of data under the control of an enterprise.



Enterprise Data is More Critical Than Ever

Enterprise data is becoming increasingly critical. This data – which can include product, customer, financial, supplier data, and more – must be available to move outside the enterprise to be used by buyers and suppliers across multiple locations, between systems in order to derive higher ROIs from existing systems and most importantly, be instantly visible.

This movement of enterprise data implies that it must be understood and employed by users and systems the way they want it, not necessarily the way it is. In order to have access to the data, there has to be a way to implement this necessity without creating additional data management and storage requirements.



Data Trends

Data Moving Outside the Enterprise

There has been an explosion in the adoption of core infrastructure enabling technologies that has enabled data to move outside the enterprise. These technologies include high-speed broadband, the World Wide Web, secure corporate intranets and portals, and networked desktops among others.

Data Used Across Multiple Geographies

Businesses with worldwide presence as well as companies selling via the Internet regularly move data across geographic boundaries. When data is transferred across these boundaries, language translation of text is only one of many transformations that may need to be applied. Numeric transformations of varying complexities are often also needed.

Data Moving Between Systems

Data now moves between systems in amounts unheard of even a few years ago. ERP systems have become de facto enterprise operating systems that must communicate to a variety of best of breed or specialty systems. These specialty systems include Customer Relationship Management (CRM), Product Life Cycle Management (PLM), Supply Chain Management (SCM), Data Warehouse, Online Analytical Processing (OLAP) and Spend Management to name a few. In addition to being consumers of data, these systems are also producers of data. The addition of each new system creates more demand for other systems to be able to use the incremental new data. This fan-out of data usage means that no single system contains all the data it uses; it is dependent on other systems to provide data.

Data Needs to Move in Real-Time

Most of the business systems described above require data in real-time, not just batch, to function at top level efficiency. Changes identified in one system need to be communicated to other systems. SCM, inventory management and purchasing systems all require real-time flow of information in order to create and maintain a smooth flow of product related information through

the enterprise. CRM, Business Intelligence (BI) and OLAP systems need real-time communication to produce time dependent customer profiling and simulation. All web service applications function at the transaction level not the batch level. In short, in the current world of business communication, there is a need for speed. We are in a real-time world.

Data Needs to Have Improved Meaning and Context

New systems create new format requirements for data. More importantly, there is a need for an improved understanding of the meaning and context of data. For example, in a CRM system, the use of a P.O. Box is acceptable for sending an invoice but unacceptable for sending a shipment. In this case the semantic use model is wrong.

As more systems use data from other systems, meaning, context and use models become increasingly important. These issues can range from cryptic abbreviations that are important in one system but unknown to other systems, incorrect use of the data (the P.O. Box example), or data that while in the correct format may convey the wrong information to the importing system. A data field header does not provide the answer. It says what the columnar information should be, but often it is not. The heading name does not ensure what the data truly is.

Impact of Enterprise Data on Core Business Drivers

Every enterprise must control and manage a set of business drivers that do not change over time or technology. There is always the need to:

- Meet competitive challenges
- Improve effectiveness of customer interaction
- Improve operation efficiencies and reduce costs

Today these business drivers are impacted by the exploding mobility of enterprise data.

Meeting competitive challenges: Companies are always racing to get new products to new markets faster. If suppliers, partners, systems and their users understand the data, new products can be designed, manufactured, distributed and sold quicker and over a wider market at a higher quality. Conversely, if the data is of poor quality or cannot be understood where and when it is needed, meeting competitive challenges is a risk area for any firm.

Improve effectiveness of customer interaction: Customer self-service via the Web is one of the most effective ways to foster customer satisfaction. Customers are able to purchase whenever they want in a routine and straightforward manner. In fact, they are demanding this purchasing experience. Businesses prefer this because the cost per transaction is low and sales via the Web are a high margin business. Furthermore, sales via the Web allow sales personnel to concentrate on their high value customers. The key to Web buying is that customers must have understandable information to know what they are buying as well as to allow for comparison with other items. This requirement indicates that the data must be presented in a parameterized or attribute specific way.

Improve operational efficiencies and improve costs: Company managers are always seeking ways to improve operational efficiencies and reduce costs. Data issues impact every facet of an enterprise when improving operational efficiencies. This is accomplished by:

- Reducing human dependencies
- Streamlining processes
- Improving reporting quality
- Improving scalability
- Supporting centralized system strategies while improving the flexibility of IT systems

Whenever data is available to a user or system the way it is needed, understood and in the timeframe desired, there is a net improvement in the operational efficiency of the enterprise.

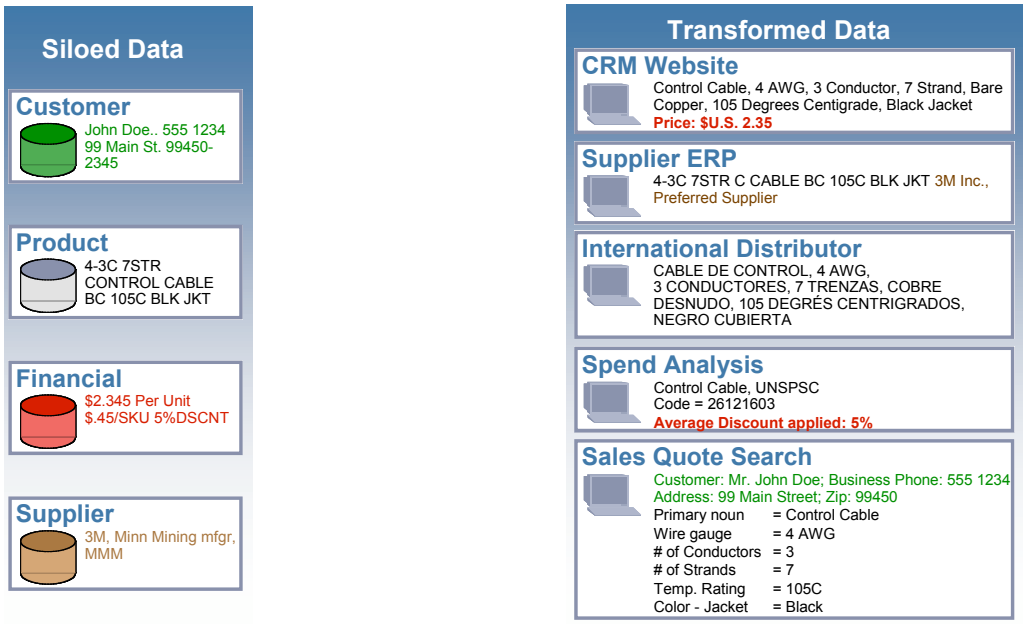
Enterprise Data: The Way It Is – Not the Way You Need It

Enterprise data often resides in the databases of the company's back office systems in its resident form, under the control of the firm's IT teams. It is typically understood by specialists but needed by generalists.

Data often needs to be combined and transformed to solve a variety of business problems. The following siloed data environments, consisting of customer, product, financial and supplier systems, are shown on the left. These data environments need to be combined and transformed to satisfy a variety of business uses. Some obvious transformations are shown below in an expanded format on the right.

- Expanded product description and associated pricing information needs to be on a website
- Product information must be sent to a supplier ERP system in a different format or standardization
- An international distributor needs product information translated into the local language for buying acceptability
- Content needs to be classified for numerous reasons from spend analysis to customer imposed categorizations
- Product data usually contains many performance, material and other attributes that define its suitability - it is necessary to be able to search for these characteristics

This content needs to be provided in real-time, as required by the growing information technology trends discussed earlier. Viewed at the record level, it is clear why simple field level mapping will never give the content user the data the way they want it. Field level or data dictionary transformations do not offer the necessary level of granularity.



Needed: Real-Time Semantic Transformation

Possible Approaches

In the example above, how can you get data from “how it is” to “how you need it”? There are a variety of ways including:

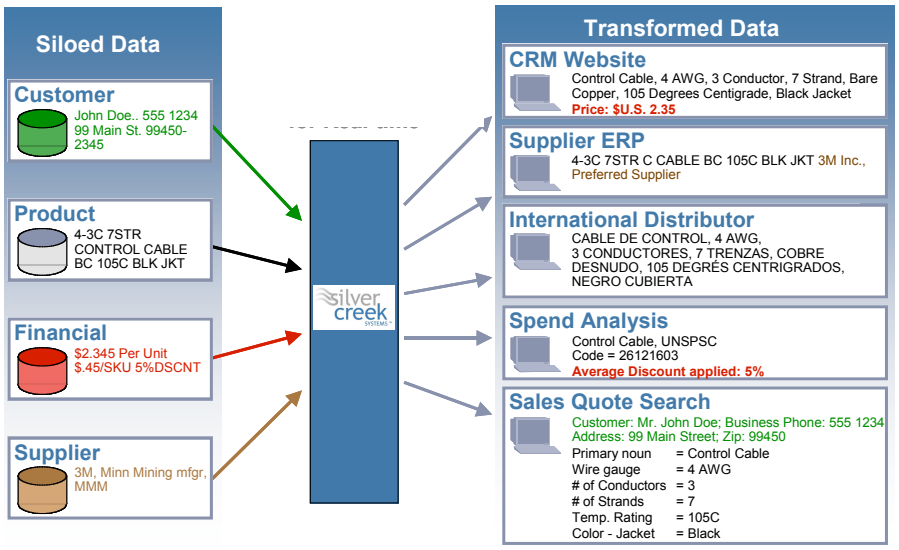
- Manual transformation and storage in specialized repositories
- Batch or bulk transformation using tools geared to handle special data domains coupled with manual transformation and storage for other data domains
- Real-time field level mapping and associations without record level transformations
- Real-time record level transformations

Manual transformation and storage: This is an appropriate solution if the number of records is no more than a few thousand and don’t change very often. If there are thousands of records that constantly change, or there are many transformations involved, then this process is unsatisfactory. A manual process does not scale, cannot respond quickly, and for this type of content is error-prone.

Mixture of Specialty Transformation Tools with Manual Transformation – There are products that effectively transform content related to simple name and address standardization but are not effective for multiple or complex domains associated with many parameters such as attributes of physical items. Manual transformations must be applied to those items in complex domains. The entire set of combined information is then stored. The problems here are not a solution that scales and one that suffers from having to synchronize the various process and data sources.

Real-time Field Level Mapping – A system that can map and aggregate information from different systems in real-time and is based on field or heading names such as “name” or “description” does exist, but it has limited if any ability to transform data at the record level. Unlike the other solutions there is no need to create mirrored repositories so data synchronization is not a problem. However, this only provides part of the needed solution.

Real-time, Record Level Transformations – Ideally, there needs to be an architectural component that resides in an enterprise’s software communication infrastructure that would transform the record data stored in repositories from multiple sources in real-time without creating another database. The data would be understandable to other systems, locales and applications and the system would understand the meaning and context of terms, abbreviations and phrases in a record at a semantic level. It would perform these transformations without the need to create mirror copies of the data and would eliminate the problems associated with data synchronization. This semantic transformation is depicted in the following figure.



Semantic Transformation Characteristics

Desirable characteristics necessary for semantic transformations require an in-place architectural "black box" acting as a transformation engine. Desirable characteristics for this engine include:

- Performs transformations at the data record level
- Performs transformations that are extremely accurate and exceptions must be identified by the system, not the user
- Performs transformations on continuous data streams in real-time
- Doesn't create additional images of the data that must be stored or synchronized with the original data generating additional management problems
- Integrates into existing business processes and data calling applications
- Rapidly adapts to changing business requirements
- Has minimal impact on IT organizations

Transformations

There are several core transformations which either individually, or in combination, allow most users and systems to receive data for maximum downstream benefit.

Standardize

Make enterprise data adhere to corporate standards

- Ensure data is understood by users, buyers, other systems and stakeholders
- Standardize data for consistent use
    - Resolve ambiguities and apply common terminology
    - Correct embedded errors, abbreviations and misspellings
    - Make enterprise data easier to read and understand
    - Enable commercial transactions to occur easily

Classify

Aggregate data ontologies and taxonomies

- Rapidly assign products to standard classification schema
- Reclassify automatically when standards change
- Publish classified products for ready retrieval
    - More ways to find similar parts

Translate & Localize

Embrace globalization by driving data across multiple geographies

- Clean, standardized data is translation-ready
- Context is understood, accuracy is assured
- Translate new parts automatically on-the-fly in real-time
- Quickly add additional languages
- Apply automation that works

Extract Attributes

Make relational databases behave properly

- Identify attribute names and values so they can be placed in databases and maintained
- Identify missing values that are not easily identified
- Enhance search capabilities with correctly stratified information

Transform Records with Multiple Columns from Multiple Sources

- Enable business processes needing multiple transformed data sources to be realized
- Enable a common understanding of disparate data sources through the adherence of internal corporate data standards

The Silver Creek Systems Data Refraction™ technology and the DataLens™ System supports these capabilities. The following section introduces the architectural components that form the DataLens System.

Silver Creek Systems Technology

The Silver Creek solution enables companies to capture and understand the true meaning of their back office content by having their subject matter or content experts (SME) capture this meaning in a way that can be reused by other users and systems both inside and outside the organization. Business analysts use the semantic information created by the SME to create a semantic map representing complex business processes that then utilize the semantic information from multiple input data streams to create simultaneous multiple output data streams as shown in the second figure.

The key to creating the semantic understanding of back office data is the creation and deployment of a metadata transformation repository that understands, at the record level, the contextual meaning of the terms and phrases of each record. The metadata information is the key that enables real-time performance without the creation of mirrored databases.

The Silver Creek Secret Sauce

Silver Creek's unique patented approach combines breakthrough technologies from multiple scientific disciplines into a modular software architecture that delivers unprecedented speed and accuracy in cleaning, classifying and translating large volumes of catalog, product and corporate data in a cost-effective way. The ingredients in our secret sauce include:

- Context comes from the data, not the data model - context is determined "bottom-up," not "top-down"
- Context is everything - contextual understanding is the key to accuracy
- Capture and reuse of subject matter expertise - people know their content, computers can't deduce it
- Small samples, tremendous efficiencies – small amounts of content mimic large data sets
- No programming or scripting – it's designed for average users, not programmers

Context Comes from the Data: There are many systems that create metadata for integrating data systems and repositories. The intent is to provide a virtual view of the data as if it were contained in a single system. These systems operate at the field or data dictionary level. The disadvantage to this approach is that the record information may not provide enough information to be complete and, in fact, may not be useful. For example, a Description field doesn't tell you what it is. An Address field containing a P.O. Box cannot accept anything other than post office mail. In the Silver Creek process, contextual information is defined at the record level, so a post office address would be known to be a semantically inappropriate shipping destination for a couch.

Context is Everything: If you were asked to interpret the abbreviation "blk" in the two phrases "blk wood" and "blk paint" you would probably say the meaning is unclear or ambiguous. "Blk" could be "black," "block," or even "bulk." This is very common with enterprise data so in order to be certain you would need to understand the context of the phrase or sentence - it's meaning in terms of other provided information. If we know the context of "blk" as it relates to wood and paint, say "block" for wood and "black" for paint, then it is readily understandable so long as we know we are describing wood or paint. Only someone who understands the context can provide the proper interpretation. Other systems often resort to statistical analysis - guessing based on how often "black" or "block" appears, but this does not work because the statistics do not have context for the solution. Just because "paint" appears more often than "wood," that doesn't mean that "blk wood" means black wood.

Another example - consider two product descriptions:

- 3 M duct tape
- 3 M Hz oscillator

Suppose you want to change one of the “3 M’s” to “3M Company.” In addition, you have a mixture of thousands of electronic parts and tape products. How are you going to differentiate the company from Mega, or Micro, or other M items? In order to interpret abbreviations, you need to understand the *context* of the phrase or sentence in which it is being used and you need to have information provided that describes the “context” of any words in question. You cannot afford to look at each part item, one at a time. Ideally you want to apply some general process that does this automatically. The Silver Creek “context engines” capture word meaning and contextual use for automated cleaning, standardization, classification and translation.

Capture and Reuse Subject Matter Expertise: Most new products are variations of existing products, and are described in terms of a set of characteristics, attributes, behaviors, states, or other information. This content tends to have “class” or type definitions and corresponding sets of attribute/value pairs. A “red fabric upholstered lounge chair” and a “blue suede leather office chair” both describe chairs in terms of the attributes of color, material, and type. If the chairs were available in 5 colors, 10 kinds of upholstery (5 leather, 5 fabric), and 10 different types of chairs, then these three sets of attributes define 500 different chairs. Furthermore, the order of the attributes in a description is irrelevant. If the first chair is described as “lounge chair upholstered in red fabric, it would be the same chair.

In other words if the attribute value pairs are defined, all permutations of the possible chairs can be automatically specified. This is what we refer to as content re-use. A subject matter expert (SME) defines the attributes and span of permitted values their content may take, along with the concepts defined above, then our system applies this throughout the data set.

The advantage of reusing a phrase or set of terms and knowing the context of its use is that we are able to apply that phrase (and other terms and phrases) to thousands records of unseen content with 100% accuracy. The result is a highly scalable and accurate system that enables the capture of SME knowledge for reference and reuse by future users. With Silver Creek, user knowledge is captured on-the-fly and immediately reused to quickly clean, classify and translate new enterprise data. Capture it once; reuse it over and over again.

Small Samples, Tremendous Efficiencies: When creating a semantic repository/knowledge base, it is necessary to represent a data set of thousands or millions of product records and a tree structure of the terms, phrases and their context. The question remains, how do we know when the knowledge base is sufficiently developed to represent the product content without evaluating each and every item manually? If the content contains 50,000 items of office supplies, what percentage of the items must be used for creating the knowledge base so the 50,000 items are likely to be represented? Silver Creek’s sampling methodology is applied to provide this answer.

We do this by first partitioning the data set into randomly sampled subsets, then creating a knowledge base for the first subset. This knowledge base is then applied to the second subset and refined based on the content of the second subset. The augmented knowledge is then applied to the third subset and so on. This process converges to a point where evaluating additional content generates little or no reusable terms and phrases, and every term is used only once. In most cases, more than 98% of the terms and phrases can be identified with small SME-processed samples. At this point we can say that the knowledge base has *converged*.

This all means Silver Creek’s sampling methods can quickly build knowledge bases and custom dictionaries in a fraction of the time it would take to clean or translate data manually.

No Programming, No Scripting: Using the technologies previously described, a knowledge creation process coupled with the sampling and other methodologies creates a very powerful system. If done traditionally, programmatic implementation and the creation of knowledge bases would be painfully slow and require each grammar rule to be written via a text editor. This would mean that specialized expertise in computational linguistics would be necessary along with computer programming skills and knowledge of statistics, sampling theory and several programming languages. The question follows – what skills would a user who is not a linguist or programmer need to possess in order to work with such a system? A solution like that would be comparable to using the Internet without the World Wide Web.

Silver Creek's solution assumes the following on the part of the users:

- Users are experts only in their data. They are SMEs or business analysts who know what data needs to be combined to solve a business problem. They know special abbreviations and acronyms relating to their data, how they want the information to be organized and structured, and are familiar with the schema or taxonomies they wish to use.
- Subject matter experts are not assumed to be technical translators. If translation of enterprise data is involved, technical translators have access to the component phrases and contextual tags associated with each phrase.
- No specialized knowledge on the part of the Silver Creek user is required. Silver Creek couples GUI operations with the context identification and grammar writing process, and then coupled with grammar parsing engines to achieve orders of magnitude increases in performance.
- The Silver Creek System is easy to use. Data owners and administrators transform data with simple drag-and-drop operations.

The Raw Ingredients

The raw ingredients of our secret sauce come from the scientific disciplines of sampling and experimental design, artificial intelligence and expert systems as well as computational linguistics.

Artificial intelligence and expert systems capture the knowledge of SMEs, the people who know their data, and apply it in an automated and reusable fashion.

Sampling and experimental design applies statistical analysis to extract small amounts of SME knowledge that represents large volumes of similar data.

Computational linguistics applies computer science technology to speech and language models derived from descriptions of enterprise data.

Summary

Breakthrough technologies that work together – our secret sauce – give Silver Creek a best of breed product for cleaning, classifying and translating enterprise data. These integrated technologies enable our products to:

- Process hundreds of thousands of records in hours, not days
- Perform unattended real-time processing
- Attach quality metrics for every line item description avoiding line-by-line quality checking, an impossibility with millions of product descriptions
- Employ a user interface that quickly and simply captures and deploys SME knowledge

Silver Creek provides innovative software addressing the complex tasks involved in making back office data available to users and systems that need this data presented the way they can easily understand it and readily use it. Sellers are able to get product data to new buyers and markets more effectively. Bill of Material, inventory tracking and distribution systems become more efficient and business intelligence and online analytics systems give clearer pictures of business operations. Silver Creek enables companies to realize the ROI from their prior IT investments.