creactives
INNOVATIVE SEMANTIC APPLICATIONS

www.creactives.com

WHITE PAPER

WHY SHOULD YOU CARE OF SEMANTICS AND MACHINE LEARNING TO IMPROVE YOUR COST MANAGEMENT?

Authors:

Adriano Garibotto - ***Chief Sales & Marketing Officer at Creactives SpA***
adriano.garibotto@creactives.com

Francesco Bellomi – ***Chief Technical Officer at Creactives SpA***
francesco.bellomi@creactives.com

ABSTRACT

Cost optimization is the critical element on Purchasing department's performance chart: it is a business-focused continuous discipline to drive spending and cost reduction, while maximizing business value.

Aggregating volume of spending, reducing the supplier base, improving the unit prices: all the main actions of a cost reduction program imply that you have clear, detailed, updated figures regarding the actual Consumption Model of your enterprise. For multinational or large companies -usually with more ERPs- this may not be an easy task. Most part of BI systems is based on the spending category manually assigned in the ERP. If the categories are few, the figures are too generic to be directly exploited, because they are a mix of very different things. That means a manual reworking to aggregate data in a more granular categorization. If the category system is deep, with hundreds of different categories, the experience said that the error that occurs in the category assignation is huge. Again, the figures coming from those reporting systems must be reworked manually to cleanse them from errors.
What really enables Cost Reduction initiatives is a very deep visibility over the consumption model based on a category system organized "by nature". That means categorizing apples with apples, probably different among them by caliber or color, but trusting that the components of the category are all apples and there is not any intruder orange within. To get that value, the information that you need is contained in the PO descriptions, but unfortunately descriptions are unstructured information unusable by the ERPs and by BI systems.

The same happens with the SKUs. Material Master Data, managed by one or more ERPs, in different languages, with different categorizations and coding system, are a mess of duplicates and incomplete information. The reasons are many: migrations, integration of multiple systems into one and negligence on the creation of new items master. Unfortunately, the main consequence is one: stocks grow.

The problem is the ERPs nature. They are transactional systems based on structured data. If you try to get the information from them, you realize that most part of the information you need is contained in the descriptions of the Purchase Orders (PO) or in the descriptions of the Material Master Data. Descriptions are unstructured information that ERPs save, but they are not able to exploit and most part of the BI systems too.

Purchasing Departments, especially in large corporations, are conscious about the value inside the unstructured, or semi-structured, information scattered inside their IT systems. The point is how to get out that value. This white paper shows how to exploit unstructured information at 100% through computational semantics.

80-90%

of all potentially usable business information may originate in unstructured form
Merrill Lynch

FACT: USERS FILL INFORMATION IN IT SYSTEMS IN MANY PERSONAL AND UNSTRUCTURED WAYS

Purchase Orders, as well as Materials Item Master, Catalogue entries and BOMs, contain unstructured descriptions. These ones typically have the following characteristics:

- They are written by different individuals, everyone following their own criteria and rarely strictly following corporate criteria.
- They are composed by sequences of acronyms, abbreviations, codes, jargon, contracted syntax, done to be able to fit in a predefined space; these descriptions are not written in a common and natural Language understandable by everybody, they require a knowledge which is in the mind of the reader to be correctly understood.
- They are highly ambiguous and the real meaning depends on the context (e.g. 80gr. can be a weight if it is associated to copy paper, but as a measure of a sandpaper it is referred to the grain type)
- Even within the same applicative domain, their meaning changes slightly in each specific setting, and evolves over the time.
- In multinational companies, they are normally written in different languages.

When you try to get reports from that mess you confirm the paradigm: garbage in = garbage out.

Since all enterprise information systems are based on structured database, the traditionally recommended actions are:

1. Code as much as possible with a single, complete and compelling methodology. By the way, please do it with a single person in a single language to avoid some of the above problems. For the rest use multipurpose codes.
2. Implement a granular category system (taxonomy) that allows you to get from the ERPs an adequate quality categorization of the products and services purchased, especially for those using multipurpose codes.

The experience shows that these measures are not effective. The reasons are many:

- Have a single coding policy strictly followed by a single team that serve all the company is probably a monstrous bottleneck that every company want to avoid.
- Coding data is too complex/too costly, especially for items that are not managed as SKUs (like most part of MRO materials) or for services.
- The completeness of the info for each item master is unsatisfactory because the coding scheme needs to be completely defined upfront, when some key information about the data may not be available.
- The nature of the data is changing over time, and keeping track of all the new kinds of entities is not feasible, especially for MRO materials that are two orders of magnitude more numerous than BOM's materials.
- The more granular is the Taxonomy, the more erroneous is the assignment of a category. This happens because the ERPs are not designed to easy search the right category. Not to mention the suggestion of the right category: in this case, we are talking about pure science fiction.

CONSEQUENCES: NO VALUE FROM THE UNSTRUCTURED AND SEMI-STRUCTURED DATA

The most common problems that Companies facing with this kind of data are:

- **Reporting activities are generic and imprecise**.
  Every time users try to aggregate, to cleanse, to normalize, to synchronize those information, it is required huge manual effort or expensive external services. Many Companies try to automatize the process by implementing BI tools. Neither the most part of BI tools solves the problem because, as we said before, they process structured data. There are some others that try to cateogorize unstructured data with "by example" training, but domain-specific data require external knowledge, that is mainly in the mind of the reader, to be properly handled. The results are unsatisfactory both for the precision of the categorization and for the percentage of categorization on the total data available. Again, manual rework is required.

- **Cost Reduction programs are less effective**.
  Large Companies, more and more, try to get competitive advantages by organizing the Purchasing Department with Category Managers for raw materials, commercial components, services, utilities and so on. To reduce costs, you will need tools that are able for example:

  - To make an internal benchmarking to discover which Company branch has better prices for a specific item;
  - To aggregate purchasing volumes to achieve incremental savings from the market.

  These tools must accurately categorize and aggregate unstructured information over domain-specific, fine-grained hierarchical taxonomies, otherwise is like comparing apples and oranges. Some SW applications available use UNSPSC taxonomy as industry standard, but is not enough to properly handle these issues and, again, manual rework is required.

- **Search activities inside IT applications are complicated**.
  As professional users, all of us have a clear experience with the searching tools inside business IT applications and how much frustrating they can be. If we use a common search by keywords, the normal results are dozens of pages to read carefully trying to find the right result, or, in alternative, full fill "advanced search" format with information that, probably, we do not know. That create a lot of problems:

  - Imagine a maintenance guy trying to find the right spare part to restart a production line: every minute is a profits loss. If the searching tool provides dozens of possible items, this means a wasting time to search the right item and, maybe, the profits loss for the interruption of the production is incomparably higher than the value of the spare part.
  - Every time you create a PO or a new item master, every ERP needs to assign a purchasing category. Large Companies usually have their own taxonomy with hundreds of categories organized in many levels. Even if you adopt standard Taxonomies like UNSPSC or eCl@ss, we are speaking about hundreds of possible categories among which choosing the right one. ERPs do not provide any aid to do that, nor control if the choice is the right. The consequences over the reporting are already explained above.

  There are some applications focus on "Natural language understanding" that try to solve real time this problem, but these approaches are not adequate: semi-structured data are not "syntactically" nor "orthographically" correct from a "Natural Language" standpoint, then the results are unsatisfactory.

SOLUTION: SEMANTIC SEARCH & CATEGORIZATION (SSC)

Let us keep it simple: imagine an Engine that has the knowledge base of hundreds of experts of specific domains (technical sectors) that speak the main European languages and allows it to understand the typical unstructured data loaded in IT silos and recognizes the products or the services from their descriptions.

What does recognize means?

- Categorizing in a specific category of a Taxonomy an object, or a service, starting from descriptions of a Purchase Order or of a Master Data, but also exploiting other available sources.
- Extracting from the description the relevant technical attributes like measures, codes, brands…

With the Semantic technology millions of objects can be recognized at glance, but is also easy to customize to specific Company knowledge, making possible to transform the categorization and identification process in an automated one, accurate, stable and repeatable over the time.

Learn More about Creactives' Technology! (Creactives' Technology)

THE TECHNOLOGICAL ADVANTAGE

Some examples can better explain the advantages of the Semantic Technology:

Reusability

- **Creactives**: the 95% of the categorization is automatic and precise, repeatable at any time with no additional cost. (higher ratio depends only on the tradeoff between the effort for "knowledge base customization" and percentage improvements of categorization)
- **Competitors:** a limit of 70% - 80% of categorization based on statistical events, but can be significantly less, even **0%** because is dependent on the quality of the unstructured data. Extensibility is very difficult for domain- and custom-specific data. Improvements are manual with extra cost for each run.

Contest interpretation

- **Creactives**: In many languages, same words in different sequence have different meanings.
(e.g. Tube Wrench ≠ Wrench for Tube)
- **Competitors**: Only key words aggregation, prepositions are ignored
(e.g. Gump **with** Gaskets **=** Gaskets **for** Pump)
- **Creactives**: metadata interpretation (e.g. :*10mm pipe* and *10mm screwdriver*; in the first case 10mm refers to the diameter of the pipe, in the second case to the length of the tip.
- **Competitors**: no way to say that the first is a diameter and the second is a length.

Independency from Language

- **Creactives**: Categorization of goods and services independently from the languages used in description
(e.g.: biro azzurra**=** ballpoint pen blue**=**Bic azul= Kugelschreiber blau=ブルー= Stylo à bille bleu)
- **Competitors**: Aggregation is done by equal or similar sequence of letters with no referring to its meaning (e.g.: above descriptions are different things)

Implicit knowledge management

- **Creactives**: Extraction of implicit information (Toner C13S050474 **=** Epson laser toner yellow)
- **Competitors**: Nothing that is not explicitly written can be categorized

ABOUT CREACTIVES SPA

Creactives has a recognized leadership position in Data Cleansing, enrichment and Analytics market. The Company is based in Verona – Italy, and has started its activity as a Software Vendor in 2008, with 50% growth from 2014 and more than 50% generated internationally. The Company has developed a prestigious client base in Italy, Spain, Germany, Belgium, France, Portugal and Sweden.

Its breakthrough Semantic Technology ensures very rapid returns for its Clients and enables them to achieve increasingly better results and higher efficiency as they continue to use the Company's products.

After being recognized by Gartner as "Cool Vendor in Advanced Data Management 2012", Creactives has been mentioned in Gartner's "Hype cycle for Analytic Applications", "Hype cycle for Procurement" and in November 2015 in the "Magic Quadrant for Master Data Management". As a very rare distinction, Gartner has cited again Creactives as "Cool Vendor in Italy" 2016.

ABOUT THE AUTHORS

ADRIANO GARIBOTTO, co-founder and Chief Sales & Marketing Officer at Creactives SpA., during the last 20 years he has accumulated a vast experience in large multinational companies, implementing Creactives' SW products focused in the Sourcing to Procure process and in Master Data Management. In many cases, he was also involved in the project deployment, supporting the data analysis, designing the cost saving solution, the implementation and the results monitoring.
MSc in Civil Engineer at Genoa University and MBA at Universidad Adolfo Ibáñez of Santiago of Chile.

FRANCESCO BELLOMI, Chief Technical Officer at Creactives SpA, has spent the last 15 years developing industrial solutions based on Machine Learning, Natural Language Processing and Knowledge Representation technologies