



## Use Case 3 in Technology Provider - Research Specialized Cost Management Services

**KW Partner:** Creative Consulting

### 1. Overview

#### **Challenge**

*Creative consulting S.p.A company have developed the system for the management and monitoring of enterprise purchase processes. The semantic based system (that will be described in the following paragraph) is composed by two specific tools: HyperCatalog and SmartSearch. They are based on a domain specific ontology, and can be used to search and buy products, and navigate the purchasing model.*

#### **Solution**

*To develop a semantic based system that semi-automatically unveils and analyzes the clients expense perimeter.*

#### **Why a Semantic solution**

*The use semantic based engine will enable to model the company organization and to relate these concepts to the purchaser officers. With a system based on ontological model, according to a structured representation of purchasable items, navigate a semantic based catalogue.*

#### **Key Business Benefits**

*Specialized in projects of costs reduction.*

#### **Business Partners**

*None.*

In this article a semantic based software system for the management and monitoring of enterprise purchase processes is described and a paradigmatic case study (the Creative Consulting S.p.A company that have developed the system) is presented. The system enables purchaser officers to search products through a semantic based engine, and navigate a semantic based catalogue in order to electronically buy the more suitable (less expensive) products. This system is based on domain-specific ontological model, developed according to a structured representation of purchasable items. In the following paragraphs some of the difficulties that has been overcome will be described. In particular the pre-analysis \_ through text mining techniques \_ of a system of documents written in natural language (that it is used to unveil concepts), and the definition of the notion of “functional equivalence” between items (that is used to effectively compare products) will be deeply analyzed.

#### **Keys components**

##### Existing Software

##### Research and development

##### Technology locks

*Pre-analysis text mining*

*Ontology development*

*Knowledge extraction*

The semantic based system (that will be described in the following paragraph) is composed by two specific tools: HyperCatalog and SmartSearch. They are based on a domain specific ontology, and can be used to search and buy products, and navigate the purchasing model.

Creative HyperCatalog manages the purchasing model, such as a data base that coherently integrates both information on catalogues, and purchasing policies (the definition of special prices or service


level agreements). As described in Figure 1, purchaser officers (or simply users) can navigate the catalogue, looking at products that wish. Selecting a category, users get automatically other sub-categories, arriving to the specific products that they need. The final proposed products will be the more suitable for users, in other words, the less expensive ones that present similar technical features.



Copyright - 2005 Creative Consulting SPA - Powered by ACP Suite - Tutti i diritti riservati



**Figure1. Creative Hyper Catalogue categories**

The choice derives from the comparison of products (described in various catalogues) according to functional and technical features, and as described in Figure 2, the purchaser officer can directly forward the order to the supplier who offers the more convenient products. Creative SmartSearch allows purchaser officers to search for a specific product using natural language. As it is depicted in Figure 3, the SmartSearch interface is similar to a common search engine, but the search mechanism (based on semantic instruments) and the quality of product identification are completely different.

[Office Supplier] 

Radice > BLOC NOTES, QUADERNI E RUBRICHE > FOGLIETTI RIPOSIZIONABILI > FOGLIETTI RIPOSIZIONABILI DI ALTRO COLORE > **FOGLIETTI DI ALTRI COLORI**

**Totale item: 2**






	Descrizione	Codice	Prezzo	Fornitore	Link
	12 CF 3M Post It note giallo, 38X51	949037.0	3,75	Office Depot - Core List	<a href="#">Info</a>
	12 CF 3M Post It note giallo, 76X76	197144.0	7,95	Office Depot - Core List	<a href="#">Info</a>

Copyright - 2005 Creative Consulting SPA - Powered by ACP Suite - Tutti i diritti riservati

Figure 2. Creative HyperCatalogue and the purchasing order request

CARTA SPECIALE PER STAMPANTI INK-JET Bianco A3


**Presente nel catalogo**

-  CARTA
-  CARTA BIANCA
-  CARTA SPECIALE PER STAMPE A COLORI
-  **[CARTA SPECIALE PER STAMPANTI INK-JET]** 

**colore:** Bianco  
**FormatoCarta:** A3

**Dettagli Ricerca:**  
 Testo: CARTA SPECIALE PER STAMPANTI INK-JET Bianco A3  
 Categoria [Presente nel catalogo]: **CARTA SPECIALE PER STAMPANTI INK-JET**  
 Attributi: FormatoCarta: **A3**  
 colore: **Bianco**  
 Tempo di ricerca: **0,25** secondi.

**Totale item: 1**

	Descrizione	Codice	Prezzo	Fornitore	Link
	RISMA 500 FOGLI CARTA RICICLATA, BRIGHT WHITE FORMATO A3 - 80 GR - COLORE BIANCO	2335545.0	1,00	Lyreco - IT	<a href="#">Info</a>

[Richiesta d'acquisto](#)

Figure 3. Creative SmartSearch

## **2. Current Practices and Technologies**

### ***2.1 Typical business practices***

The main idea behind these applications is to use text-mining techniques to build a structured representation of purchasing model, starting from items, their natural language and textual descriptions found in producers' catalogues. The structured representation is defined by an ontological model of the items' domain, which describes the taxonomical organization of the catalogue, and specifies and constrains the technical attributes of the items themselves. Besides, the natural language queries performed by the user are translated into the same structured representation. The main reasoning service enabled by the ontological model is the ability to decide whether two items are "functionally equivalent" with respect to the use intended by the purchaser; in most cases, this can be modelled by taking into consideration only some relevant attributes, while disregarding the others (as an example, the kind of tip and the length of the blade are relevant attributes for a screwdriver, whereas the color of the handle is not).

### ***2.2 Toward Ontology Driven Text Mining***

Since the acquisition and pre-processing of producer catalogs is by far the most time-consuming activity for the development of the purchase model, we are interested in providing methodologies and tools to automate these processes, while preserving accuracy. This involves acquiring and cleansing data from multiple catalogs, written in different formats by different producers, which change through time and purchaser location.

The main purpose of the data gathering and cleansing phases is the identification of functionally equivalent items along different catalogs. The purchase model details the specific policies that prescribe the choice of the most convenient producer when two or more functionally equivalent (or even equals) items are listed in more than one catalog.

The identification (and aggregation) step is not trivial, since the primary source of information about each purchasable artifact consists in a natural language textual description of the item itself, a description written by a human being for another human being, thus usually incomplete and context-dependent, potentially ambiguous, generally non providing any formally shared identification token (since identification codes are often unique only within a single catalog, or a single producer), and based on an open-ended vocabulary. Our approach consists of developing an ontological model of the target domain, with support for the notion of functional equivalence (which, in turn, is strictly domain- and context-dependant), and then, in populating a knowledge base of the purchase history, based on the schema provided by the model, extracting data from the purchase orders using specialized model-aware text-mining tools.

Since it did not seem possible to effectively address the issues related to item description classification based upon unsupervised learning algorithms, we decided to approach the problem differently.

The text mining tool is configured by "decorating" the ontological schema (consisting in entities, attributes and relations, organized mainly by hierarchical subsumption) with collections of weighted rules that recognize user-defined terminological and linguistic features, which are expected to be relevant in the source text. The rules are exploited by the three components of the text-processing tool (the classifier, the attribute extractor and the attribute normalizer), that perform a shallow parsing of the item description, in order to provide a set of tentative representations of the described artifact in terms of the ontological schema. In the end, the best description is chosen by evaluating the weight of the triggered rules.

The end user is not asked to produce example documents, but to list specific linguistic *features* that are supposed to characterize to a good extent the documents belonging to each of the categories of the taxonomy.

A feature is intended to be a word, or a phrase, or a set of words expected to occur in a limited range of word positions in the document, or, more generally, a sentence belonging to the language generated by a context-free production. A graphical tool has been developed, that allows the users to visually build, and edit context-free grammars, using a graphical version of the BNF notation (Figure 4).

The user assigns a weight to each feature, and can compose more complex features from simpler ones using Boolean operators, and other modifiers that further control the weights (Figure 5). A document is classified as an instance of each given category if the sum of the weights of the matched features meets a category-specific threshold (which is also user-defined). Negative weights are allowed, and it is also possible to assign a weight to the event that a feature is *not* found.

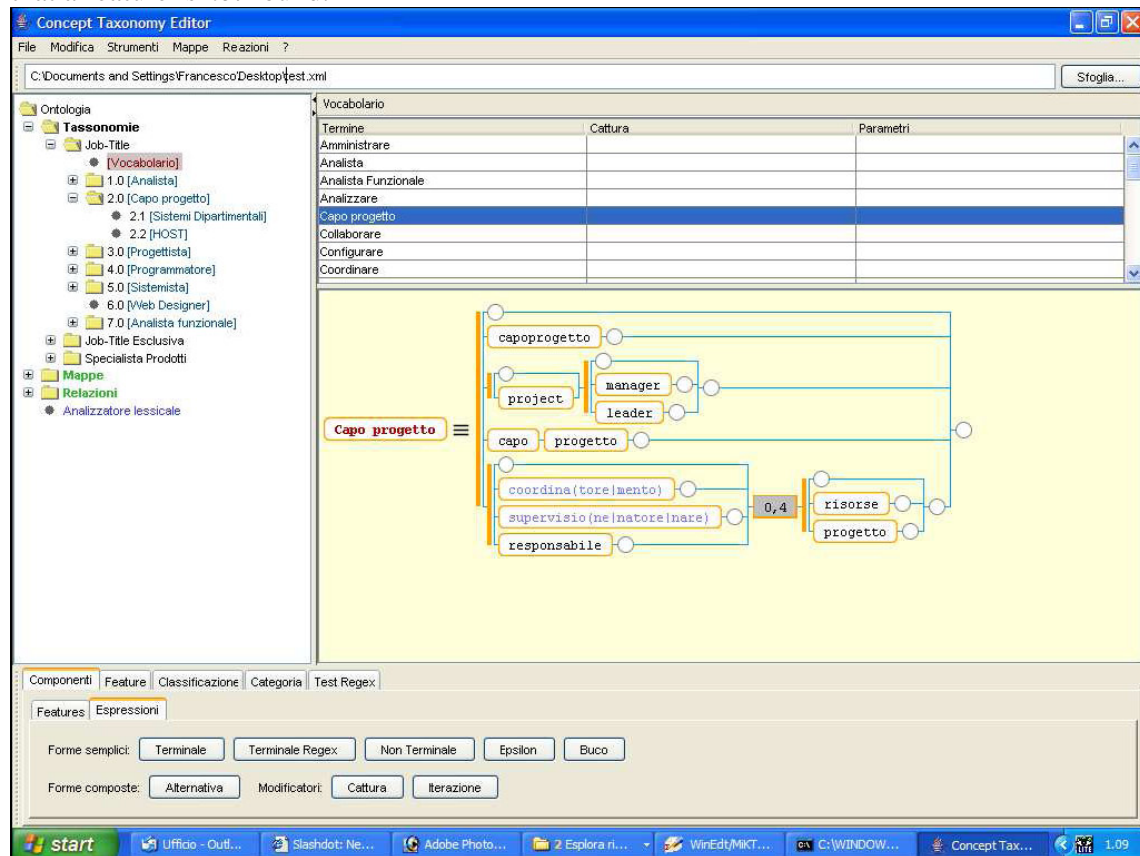


Figure 4. Visual tool for pseudo-BNF notation

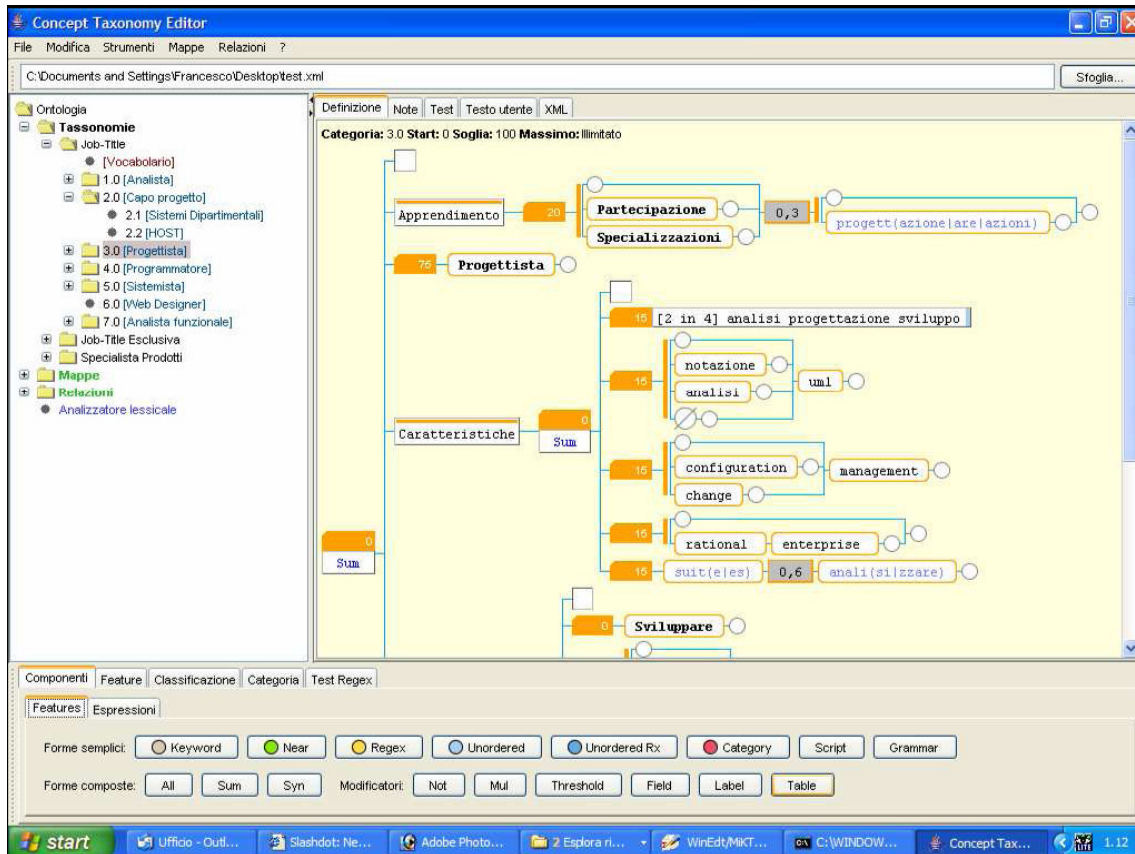


Figure 5. Definition of a category classifier by means of weighted features

Working with domain experts and end users, with the goal of defining a taxonomy and devising a suitable set of linguistic features for training, the classifier become often challenging when these people don't have some previous experience with some sort of formal taxonomical classification. Namely, we have repeatedly encountered the following issues:

- even if taxonomical classifications appear to be ubiquitous in everyday life, and people nearly always have a first-hand experience in organizing information in a hierarchical fashion (such as in nested folders within the file system of a personal computer), it is often difficult for the layman to think of an hierarchical taxonomic structure in terms of an is-a relationship; furthermore, discriminating linguistic features of a category are not, generally, inherited by its sub-categories, since those features are not attributes of the category, but rather of the category, and the target corpus combined. As an example, if we consider a taxonomy of job titles, the set of features needed to classify job applications is very different from the set of features needed to classify job offerings. Stronger feature inheritance is more likely to happen in the lower, more specific level of the taxonomy.

- sometimes, users find it difficult or unnatural to assign weights to the features; conversely, sometimes they try to be as precise as possible in the fine-tuning of the weights, in fact overestimating the sensitiveness of tools. A good practice is to ask the users to partition the features in a small number of equally weighted classes for each category.

- it is really difficult to evaluate the performance of a classifier [Basili, Moschitti and Pazienza, 2001], because most of the (niche) domains we take into consideration do not have a standardized taxonomy, or the standardized taxonomy does not fit the intended use, and, as a consequence, it is not easy to find or assembly a normative benchmark. In our experience, the domain experts are usually only able to evaluate each single judgment of the classifier,

and a failure to correctly classify often reveals an inadequacy of the taxonomical organization rather than a flaw in the training of the algorithm.

So, the following taxonomy development process has been implemented: there is a bootstrap phase, where a domain expert provides an initial “seed” taxonomy, and a (possibly large) corpus of unclassified, yet domain-related documents. After that, the corpus is statistically analyzed, and a list of relevant keywords is generated. This list could suggest some revision to the seed taxonomy, and, more important, should provide some guidance for the definition of the first version of the taxonomy annotated with the linguistic features to be used by the classifier. Then, we enter in the cyclic refinement phase, where the annotated taxonomy is used by the classifier to generate a classified corpus; the classified corpus is statistically analyzed in order to provide a more accurate set of suggested linguistic features, that should be used to improve both the structure and the annotation of the taxonomy, and so on until the user is satisfied by the taxonomy and the automated classification (for a more in depth methodological analysis see [Cristani Cuel, 2004a, 2004b]).

We are not only interested in the output of the classification and normalization process, but also in the domain model and classifier themselves, which became reusable “as is” in the same context and with similar input data, and can be used as a basis for deriving similar models for “contiguous” contexts. A simple example set is too loosely structured to be effectively usable as reliable domain model, for other uses than the basic classification process.

In the development of the above mentioned software tools, it has been found that it is of paramount importance to enable the domain experts involved in the definition of the taxonomy to have a direct and unmediated role in the development of the ontological model, even when these people did not have any previous experience in the definition of taxonomies.

### ***2.3 Actions should be carried on, aiming at exploring “architectural” issues***

The activity of Creative Consulting S.p.A. is not finished yet. Some other actions should be carried on, aiming at exploring “architectural” issues of the systems such as the sustainability of larger domain models. In particular it will be investigated:

- the optimization of development times for multi-language classifiers (using heuristics to analyze multi-language catalogs in order to suggest relevant candidate linguistic features to domain experts);
- the definition of explicit performance metrics to evaluate accuracy and discuss quality issues with customers in a more quantitative way;
- the “refactoring” of some linguistic knowledge developed for some specific domains, which turns out to be re-usable across different (and/or more general) domains.

Some other future works that deal with organizational aspects will be:

- the analysis of the type of industries (pharmaceutical, healthcare, automotive, logistics, etc.) and organizational assets (small, medium or large enterprises) that will benefit from these solutions;
- a more in-depth analysis of the co-determination between technologies and organizational assets. In particular a very specific analysis should be done, in order to study on how HyperCatalog and SmartSearch can effectively be implemented within the firm, and how this will affect to its traditional organizational processes;
- cost analysis on ontology creation. A quantitative analysis on how an ontology based systems affects the existing infrastructure is required. In particular this requires means to monitor the quality of the ontology development and deployment processes, to estimate and control the costs involved in the development and usage of ontologies and to investigate the

costs and benefits of applying particular development or deployment strategies. A qualitative analysis of existing ontologies and ontology engineering methodologies, methods and tools is needed. In particular the dissemination of ontology-based technologies at corporate level requires methods to measure the usability of a particular ontology in a specific business scenario, but also objective means to compare among methodologies, methods and tools dealing with them.

## References

Aizawa, A. (2001) 'Linguistic techniques to improve the performance of automatic text categorization'. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pages 307–314, Tokyo, JP.

Allen, J. (1995) *Natural Language Understanding*, Second Edition. The Benjamin/Cummings Publishing Company, Inc., Redwood City, California, USA.

Apte, C., Damerau, F.J. and Weiss, S.M. (1994) 'Automated learning of decision rules for text categorization'. In *ACM Transactions on Information Systems*, 12(3):233–251.

Ariba, (2005). *Ariba Web Site* <http://www.ariba.com>

Ashby, W.R., (1956) *An Introduction to Cybernetics*, John Wiley & Sons, New York.

Basili, R., Moschitti, A. and Pazienza, M.T. (2001) 'NLP-driven IR: Evaluating performances over a text classification task'. In Bernhard Nebel, editor, *Proceeding of IJCAI-01, 17<sup>th</sup> International Joint Conference on Artificial Intelligence*, pages 1286–1291, Seattle, US.

Boland, R.J., & Tenkasi, R.V. (1995). Perspective Making and Perspective Taking in Communities of Knowing. *Organization Science*, 6(4), 350–372, 1995.

Bowker, G. & Star, S.L. (2000). *Sorting Things Out: Classification and its Consequences*. MIT Press.

Chai, K.M., Ng, H.T. and Chieu, H.L. (2002) 'Bayesian online classifiers for text classification and filtering'. *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval*, pages 97–104, Tampere, FI. ACM Press, New York, US.

Creactive 2005. *Creactive Consulting S.p.A Web Site* <http://www.creative-consulting.com/>

Cristani M. & Cuel R., (2004a) "A comprehensive guideline for building a domain ontology from scratch". In proceeding of "*International Conference on Knowledge Management (IKNOW'04)*", Graz, Austria

Cristani M. & Cuel R., (2004b) "Methodologies for the Semantic Web: state-of-the-art of ontology methodology". *Column of SIGSEMIS Bulletin. Theme "SW Challenges for KM"* V. 1 I. 2

Euzenat, J. , Pin and, J.-E., Ronchaud, R. Research Challenger and Perspectives of the Semantic Web. *Strategic Research Workshop*. France, 2002.

Fellbaum, C, ed. (1998) *WordNet: an Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, US.



Joachims, T. (1998) 'Text categorization with support vector machines: learning with many relevant features'. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Klavans, J. and Resnik, P. (1996) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. The MIT Press, Cambridge, Massachusetts, US.

Maturana H.R. and Varala F.J. (1980) *Autopoiesis and Cognition: The Realization of the Living* Dordrecht: D. Reidel.

Manning, C.D. and Schütze, H. (2000) *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, US.

Mitchell, T.M. (1997) *Machine Learning*. McGraw-Hill, New York, NY, US.

Nigam, K., Lafferty, J. and McCallum, A. (1999) 'Using maximum entropy for text classification'. In *Proceedings of IJCAI-99, 16th International Joint Conference on Artificial Intelligence Workshop on Machine Learning for Information Filtering*, pp. 61-67.