Dynamic Concept-Based Taxonomy used for image recovery based on their textual description

Jaime Lara, María de la Concepción Pérez de Celis, David Pinto

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
Puebla, México.
{jlara, cperezdecelis, dpinto}@cs.buap.mx

Abstract. In this paper, we will describe a methodology for the development of a search system based on the usage of textual descriptions in order to recover images using a dynamic taxonomy. This taxonomy is stored in a relational database that is part of a system that allows the user to explore and refine his search, using a navigation tree. We propose the usage of information recovery techniques to extract a controlled vocabulary and defining concepts that will be structured in hierarchies with the aid of a thesaurus, in order to automatically generate facets. Such facets will then be linked with the studied objects using indexes. We have applied this model on a collection of artworks, linking an image with a textual description in the Spanish language. The experimental results show the advantages of such a model.

Keywords: automatic generation of taxonomies, dynamic taxonomies, faceted classification, faceted taxonomies, information recovery, navigation trees.

1 Introduction

An alternative to classical search engines (keyword-based search) is to allow the user to navigate through contents. This process of navigation requires the user to know the way in which the information is organized. A practical solution to this inconvenient is the use of a pre-established taxonomy. We must point out that taxonomies require the organization of contents around a certain knowledge domain. From our point of view, a taxonomy improves the recovery of information on specific topics, but it is still a rigid solution since it does not allow for the possibility of restructuring the users searches as he advances in the recovery of results. As such, the user cannot connect the recovered objects with other knowledge domains within such objects may be connected.

A possible solution that allows the user to be guided and also by self may interact and refine his search is the usage of what is known as a faceted taxonomy. Such taxonomy allows the information to be structured and accessed in more than one dimension. This benefits the user because he can easily and intuitively localize and explore the information using the different approaches provided by the taxonomy. In this work we will focus on the methodology used to build a taxonomy which will be

the base for the implementation of a system of faceted classification, which will then allow the user to build his own knowledge map. In the following sections we will present some of the work linked to the design and construction of taxonomies and we will analyze its usage in the recovery of information. Later on, we will discuss the methodology used for the confirmation of the taxonomy, with emphasis on the recovery of objects based on their textual descriptions. We will present the results we have obtained and we will emphasize the importance of conceptualizing the operations that allow for the creation of open taxonomies.

2 Related Works

An information retrieval system in which one has a collection of objects that have diverse properties, the relevance of such properties varies depending on the user. As a matter of fact, a solution for finding information regarding a particular topic consists in asking an expert on the subject. Due to the fact that, generally, this is not possible, one has to recur to the use of taxonomies.

The importance of taxonomies lies in the possibility of them being used as a triplet (classification scheme, semantic interpretation, knowledge map [1]) in the organization and recovery of collected objects (possibly in a database).

Taxonomies can be organized under different structures: lists, hierarchies, poli-hierarchies, multi-dimensional matrix and facets, among others. Among them, poli-hierarchies, multi-dimensional matrix and facets are linked to the possibilities of objects belonging to more than one category.

Characterizing objects under diverse categories or characteristics allows us to create a metadata model were the objects and their characteristics take on a new meaning. It is because of such a phenomenon that faceted taxonomies can take advantage of the way in which metadata behave. Metadata models are a collection of information built on some type of object, or on a part of such an object. For example, the name (or title) of an artwork, its author, date, image, textual description (denotation), interpretation (connotation) or genre, represent metadata that can be associated to a cultural object.

As such, every element in the metadata scheme can be incorporated as a concept of the faceted taxonomy, and thus it can be recovered using a search engine. Therefore it is possible to access an object under any of the dimensions under which it has been classified. If we consider the previous example, we could recover a cultural object by its genre, connotation or denotation or we could navigate through the different facets, taking into account that they are orthogonal among themselves.

Sacco [2], [3] introduced the concept of dynamic taxonomies and the notion that it could withstand the incorporation of facets that, in themselves, require an independent taxonomy for their description. Due to the fact that our objective is the recovery of images based on their textual description, we will use the fundamentals of dynamic taxonomies but we will also extend the domain of the data modeling created by Sacco, because we will include textual objects.

In the literature we can find two interesting approaches for the management of textual objects. The first such approach assumes the existence of various topics over

which one wishes to generate taxonomy, and because of that the algorithms are designed to extract the terms and concepts linked to such topics from the documents. The MindMap system [4], in particular, proposes the generation of multiple taxonomies for any collection of documents, each one with unique topics. This multiple taxonomies are visualized in the system as an integrated tool, thus obtaining a system that allows the organization of information in multiple ways. In this system, the classification of a document under any taxonomy depends on its similarities with other documents. For this purpose, a system of spatial coordinates is used, in which the similarities are determined by the proximity between the coordinates of each document. As we have pointed out before, in this approach each one of the multiple taxonomies requires a series of keywords, associated to the concepts under which the classification of the documents will be structured, in order for the analysis to begin.

In contrast, the second approach is centered in determining the different facets and defining their taxonomy from the textual analysis of the collected documents by using text analysis. A good example is the algorithm used by the Flamenco project [5] of Berkley University for the automatic generation of facets using WordNet over a corpus written in English.

3 Our approach

During the development of the information system for the management of cultural objects, we have implemented the metadata model proposed by CCO [6]. However, when extending this model with the metadata of genre, connotation and denotation, the inclusion of a textual description of the objects was necessary. This fact determines the necessity of a methodology that allows for the generation of a taxonomic structure of the facets, under which the different terms included in the description of the objects will be classified. We must point out that such a methodology can be extended to any corpus of descriptive documents over which one means to develop a taxonomy. Our proposal consists in using techniques and algorithms employed in information retrieval to generate a faceted classification, which allows for the association of part of a document to the different facets. Such facets will be generated from a thesaurus linked with a “controlled vocabulary”. This vocabulary, in turn, will be extracted fro the processing of different texts from the target collection. This process is described in detail in the following sections.

3.1 Approach by dynamic taxonomy

The largest difference between a conventional taxonomy and a dynamic taxonomy is that the former are monodimensional (one element is classified under one concept), while the former are multidimensional (Fig. 1). The term *dynamic* reflects the ability of the taxonomy to adapt to different approaches, perspectives and interests.

Sacco [2], [3] established the following inference rule: *Two concepts A and B are related iff there is at least one item D in the infobase which is classified at the same time under A (or under one of A’s descendants) and under B (or under one of B’s descendants).*

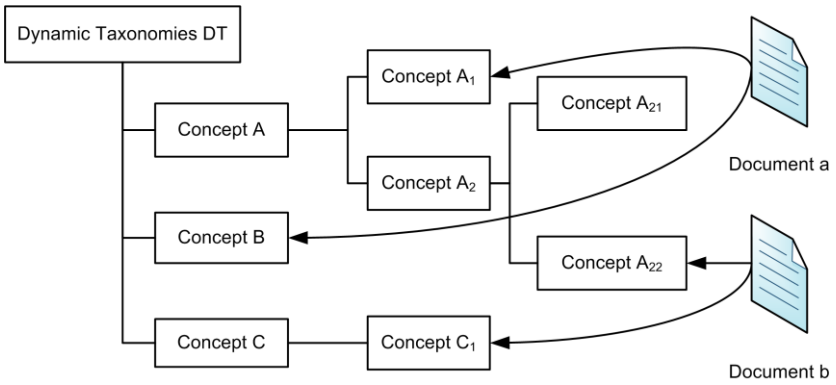


Fig. 1. Multidimensionality in dynamic taxonomies.

The left side of Fig. 2 shows a set of data classified under two facets, and in the right side can be appreciated the change that the taxonomies of our facets go through when we focus our search on the concepts B and H.

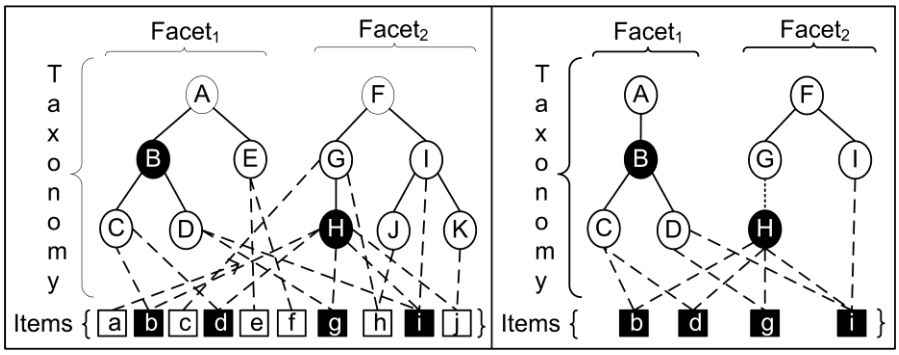
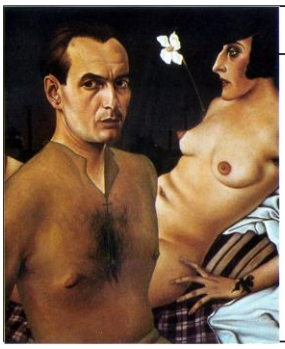


Fig. 2. Left side: Data set {a,b,…,j} classified under two facets. Right side: Reduced taxonomy.

4 Study case

The data set that will be studied consists of a total of 500 artworks [7], [8] (An example can be appreciated on Fig. 3) with description in the Spanish language, of which a training set of 200 artworks was selected for testing.

In order to implement the dynamic taxonomy we used the following facets: genre, connotation and denotation. Genre refers to each one of the different categories or classes under which the artworks can be classified according to common forms and contents. For example: portraits, self-portraits, landscapes, religion, mythology etc. Of these three facets, only genre is defined by an expert user, while the other two are obtained automatically.



Autorretrato con modelo

En este estudio de la **vanidad** decadente, el artista pinta su autorretrato, mientras una **mujer desnuda** yace provocadora detrás de él, físicamente cercana pero emocionalmente distante. El pintor viste una prenda transparente que parece una capa de piel, unida mediante puntadas en el pecho. La mejilla de la **mujer** está surcada por una cicatriz impropia. Detrás de ella se alza una **flor** fálica, un simbólico **narciso**, y más allá de ambos podemos entrever los tejados de la sombría **ciudad** industrial.

Fig. 3. An artistic object with their textual description (denotation).

5 Methodology

In the following paragraphs we will show the steps taken in order to generate a classification system based on dynamic taxonomies.

5.1 Obtaining a Controlled Vocabulary

A Controlled Vocabulary (CV) is an organized lists of words and phrases that are used to initially tag content, and then to find it through navigation or search.

5.1.1 Elimination of stopwords

The first step consists in the elimination of stopwords, since such words do not allow for discrimination of relevant attributes of the objects. There are several lists of stopwords in Spanish (Snowball[1]; Ranks[2]).

5.1.2 Stemmer

In order to obtain a dynamic taxonomy, one must have a controlled vocabulary. In this particular case we applied a Spanish stemmer[3] based on Porter's algorithm, which allows for the decrement of the controlled vocabulary and augments the definition of each concept, providing a better recall.

5.1.3 N-Grams

With the objective of finding concepts formed by more than a word, we obtained bigrams, trigrams and 4-grams of the textual description of artworks. And also, we have a set of words (unigrams).

[1] Snowball, *Spanish stop word list*, http://snowball.tartarus.org/algorithms/spanish/stop.txt
[2] Ranks NL, *Spanish stopwords*, http://www.ranks.nl/stopwords/spanish.html
[3] Snowball, *Spanish stemming algorithm*, http://snowball.tartarus.org/algorithms/spanish/stemmer.html.

5.2 Defining the Concepts

A concept is a label which identifies a set of documents[4] (classified under that concept), every concept is related with a certain level of abstraction that depends with the level in the taxonomy, this make clear that concepts are not terms. So, the problem here is how we can know what terms of phrases of our controlled vocabulary can be a concept, to answer that question we use a thesaurus, using a thesaurus we can define the part of our controlled vocabulary that we can use as concepts as we can see in the Fig. 4 and in the equation 1.

$$Concepts = Thesaurus \cap CV \quad . \tag{1}$$

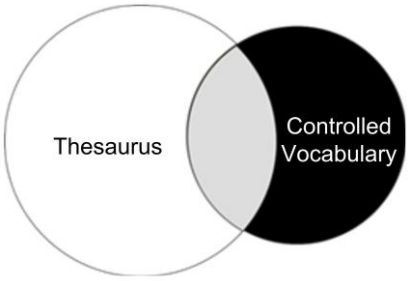


Fig. 4. A Concept definition

5.2.1 Incrementing and Expanding Concepts

By using clustering algorithms, we look for terms that were not originally considered part of the concepts using the remaining concepts of the thesaurus. Afterwards, new concepts are added using a supervised process like the one shown in Fig. 5. Also, there is a possibility to expand the definition of each concept, this leads us to the generation of new clusters between the controlled vocabulary and the concepts.

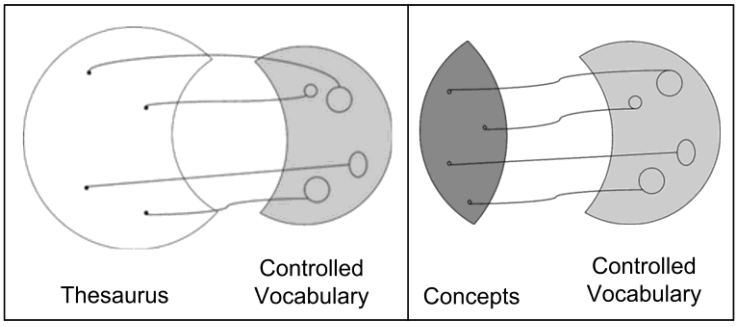


Fig. 5. Left side: Expanding the concepts. Right side: Concept expansion.

[4] Information Management Unit, *Guided Interactive Discovery of e-Government Services*, http://www.imu.iccs.gr/sweg/presentations/Giovanni%20Maria%20Sacco.ppt

5.3 Obtaining taxonomies

The next step consists of generating the taxonomies of the concepts using the thesaurus hierarchies.

5.3.1 Defining the Taxonomy

In this step we obtain the hierarchical structure of every concept based on the thesaurus structure, we use a hash table to do this process (See Fig. 6). The taxonomy is then created by the process of linking all the concepts.

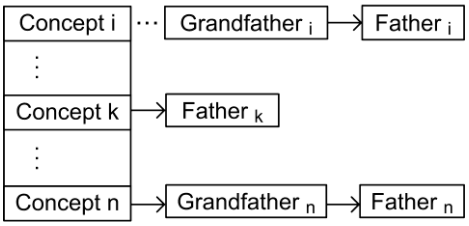


Fig. 6. Obtaining the hierarchy of every concept.

5.3.2 Structuring the Facets

Once the hierarchic structure of the concepts contained in the vocabulary is ready, one must supervise under which facet they will be placed. In Fig 7 we show a simple example of the taxonomy for the facets.

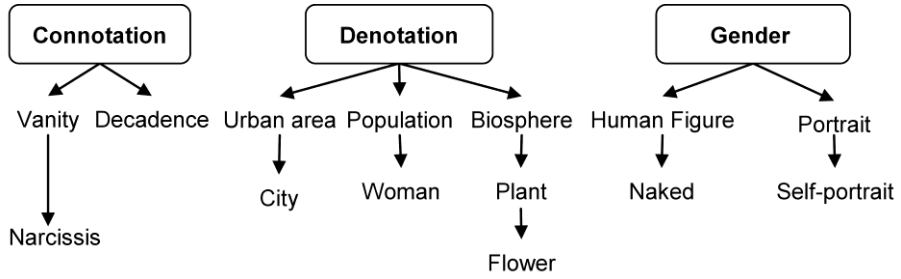


Fig. 7. A facet taxonomy example.

5.3.3 Pruning

The taxonomy should be pruning, to avoid that the user spend time expanding unnecessary nodes, this nodes are those that doesn't helps us in the filter process.

5.3.4 Frequency filter

If we order the controlled vocabulary based on its frequency, it is possible to eliminate the less frequent concepts, due to the fact that they will generally not be used for information recovery. However, our proposal consists in implementing this filter directly over the taxonomy. This process entails a second pruning over the faceted taxonomies.

5.4 Indexing

Each artwork contains a textual description. Such description has words or phrases that are included in the controlled vocabulary. At first, and before we have used hierarchies to bond the controlled vocabulary to the faceted taxonomy, the indexes are not related among each other (Fig. 8).

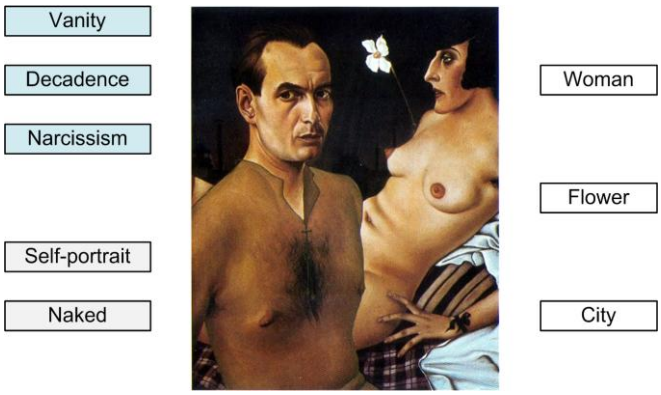


Fig. 8. Denotation index.

However, once created the taxonomies, each index is related to each other by a hierarchical scheme, as shown in Fig. 9. This bonding allows to index the artwork under the information that contains its textual description, and to add the hierarchical information of each concept (Fig. 10).

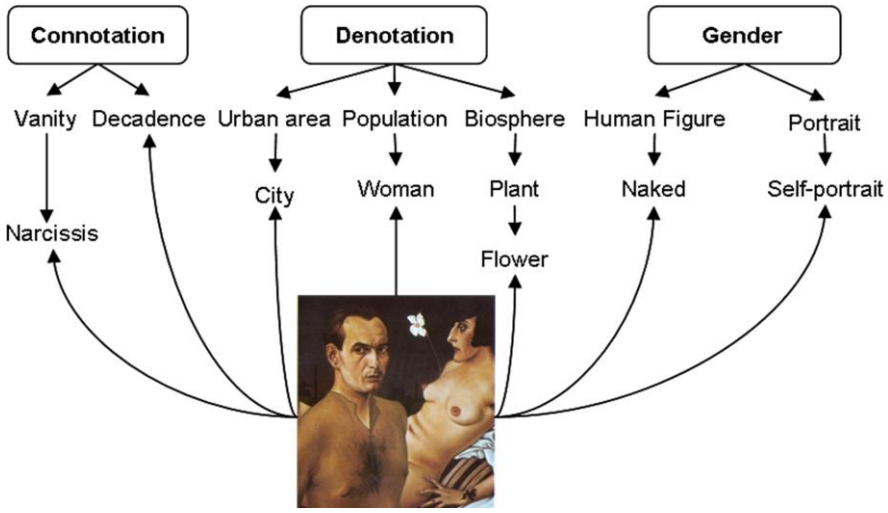


Fig. 9. Concepts connections.

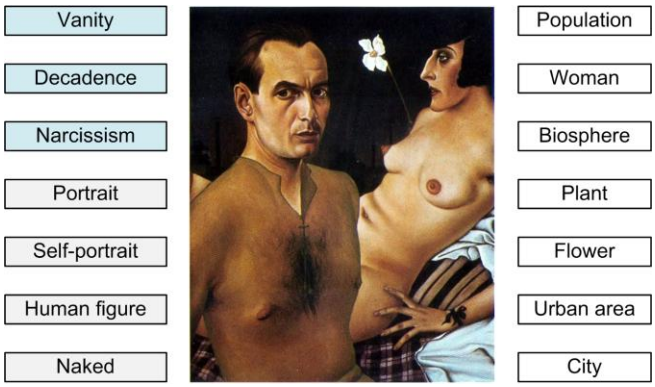


Fig. 10. Conceptual index.

5.5 Storing model

In order to use the model, the facet taxonomy and the objects should be stored, we use the Extended Entity-Relationship diagram shown in Figure 11 for this propose.

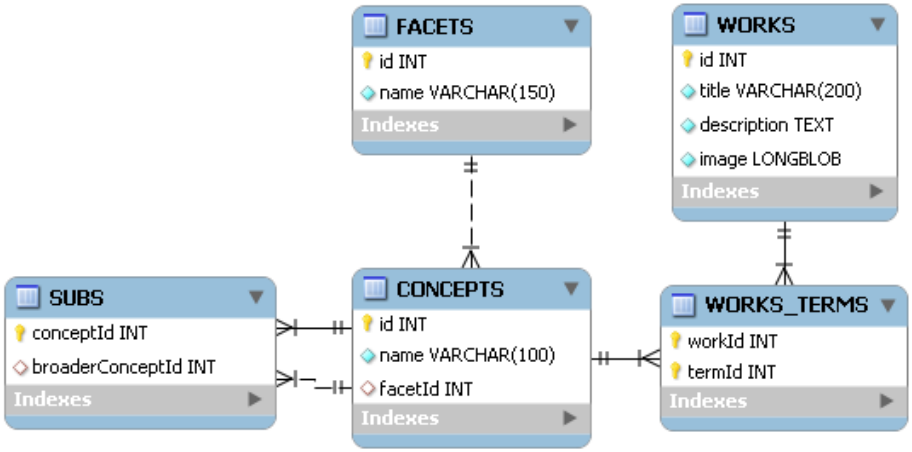


Fig. 11. EER Model for store a facet taxonomy and artworks objects**.**

5.6 Implementation (Navigation Tree)

The final step is make a visual framework in which the user can select and combine appropriate concepts [2], this is develop by a Navigation Tree [9] (taxonomic tree [1]). The navigation tree contains nodes that enable the user to start browsing in one facet and then cross to another, and so on, until reaching the desired level of specificity [9].

6. Results and Evaluation

By using the procedure described earlier we obtained a controlled vocabulary of 500 concepts that describe the 200 documents. It is important to mention that the eliminated stopwords represented 50% of the total words. By means of this procedure we were able to bond each artwork to an average of fourteen indexes.

We performed a comparison between our faceted classification system and the "Full Text Search" function of MySQL, which uses a Boolean search for any given data set. We configured a set of supervised consults, taking into account two criteria: recall and precision. Fig. 12 shows the results.

As one can see in Fig. 12, the faceted classification system provides a significantly higher degree of recall when compared to the Boolean search, underlining the advantage of using a conceptual search instead of a textual search.

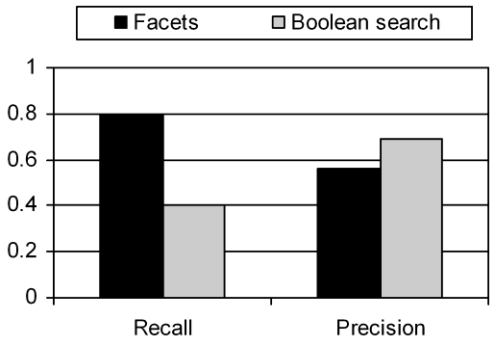


Fig. 12 Recall and Precision evaluation.

7. Conclusions

So far, the automatic solutions for hierarchy constructions available have not offered satisfactory outcomes when faced with the construction of taxonomies. However, some methodologies developed for the construction of facets and the generation of taxonomies based on textual analysis has yielded encouraging results. As of now we are implementing new algorithms that will allow the generation of interpretations based on generic descriptions, expanding the "connotation" facet described in this paper. We have also considered the usage of a terminological conceptual thesaurus for this task, and the possibility of allowing the user to use "open taxonomies" thus allowing the taxonomy to evolve, reflecting the progression of the words and their interpretation and keeping its novelty.

References

1. Patrick Lambe, Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness, ISBN: 9781843342274, Chandos Publishing (Oxford) Limited. UK, 2007.
2. G.M. Sacco, Dynamic Taxonomies: A Model for Large Information Bases. IEEE Transactions on Knowledge and Data Engineering 12, 2, pp. 468-479, May 2000.
3. G.M. Sacco, Some Research Results in Dynamic Taxonomy and Faceted Search Systems. SIGIR'2006 Workshop on Faceted Search, August 2006 Seattle, WA, USA.
4. Spangler, S. Kreulen, J.T. Lessler, J. "MindMap: utilizing multiple taxonomies and visualization to understand a document collection", . Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002. HICSS. Pags. 1170-1179.
5. E. Stoica and M. Hearst, "Demonstration: Using WordNet to Build Hierarchical Facet Categories". The ACM SIGIR Workshop on Faceted Search, August, 2006
6. Categories for the Description of Works of Art (CDWA), editado por Murtha Baca and Patricia Harpring, http://www.getty.edu/research/conducting_research/standards/cdwa/index.htmll. 2009.
7. El ABC del Arte del siglo XX, Primera edición en español 1999, Editorial Phaidon Press Limited.
8. Masdearte.com, Portal de arte contemporáneo, http://www.masdearte.com/item_critica.cfm?id=315. 2009.
9. Y. Tzitzikas, A. Analyti, N. Spyratos and P. Constantopoulos, An Algebra for Specifying Valid Compound Terms in Faceted Taxonomies, Journal on Data and Knowledge Engineering (DKE), 62(1), 2007.