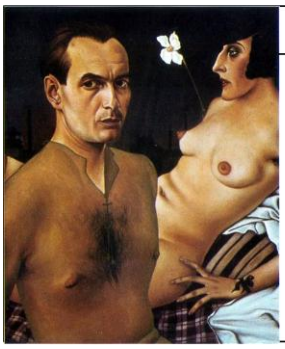

Autorretrato con modelo

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.

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

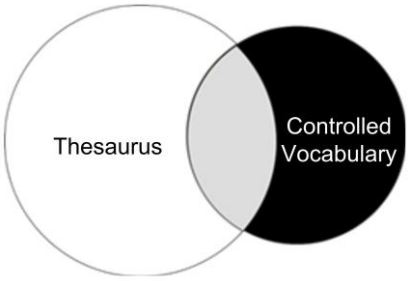

Fig. 4. A Concept definition

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.

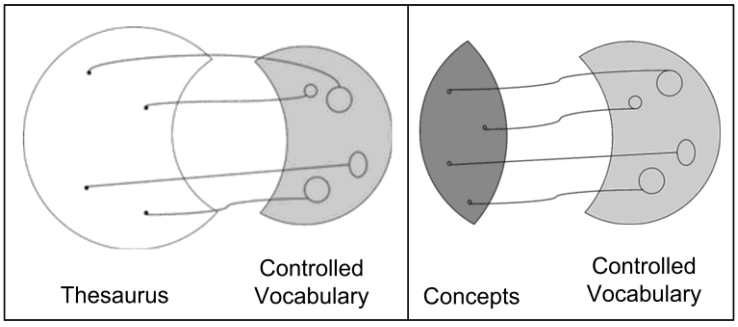

Fig. 5. Left side: Expanding the concepts. Right side: Concept expansion.

[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.

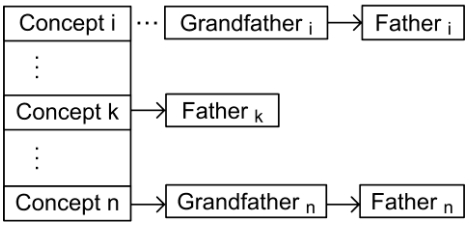

Fig. 6. Obtaining the hierarchy of every concept.

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.

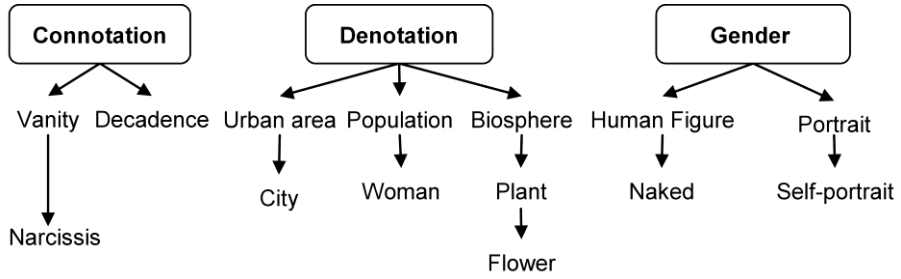

Fig. 7. A facet taxonomy example.

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).

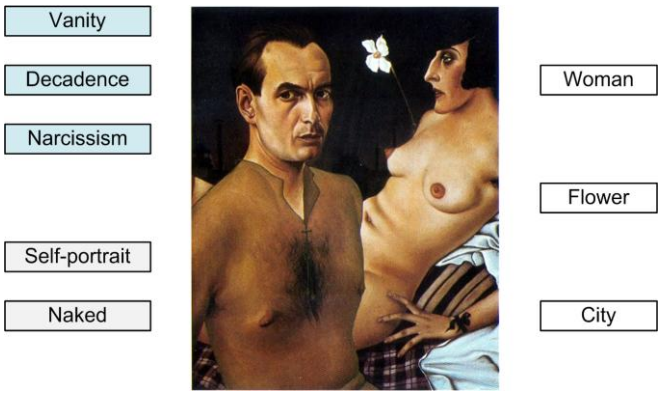

Fig. 8. Denotation index.

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).

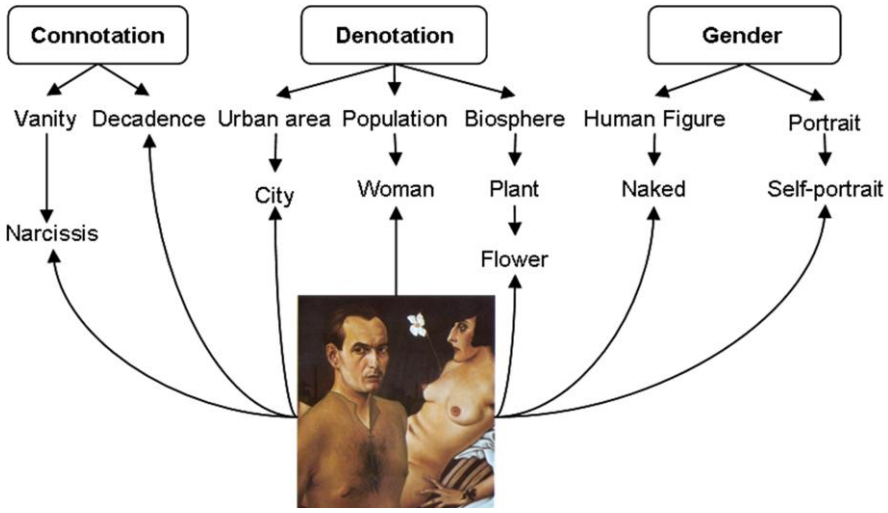

Fig. 9. Concepts connections.

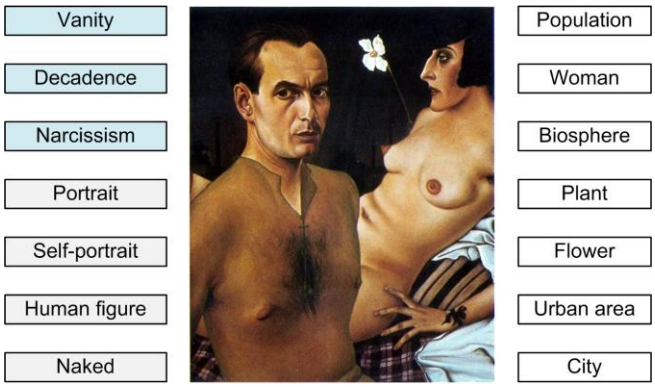

Fig. 10. Conceptual index.

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.

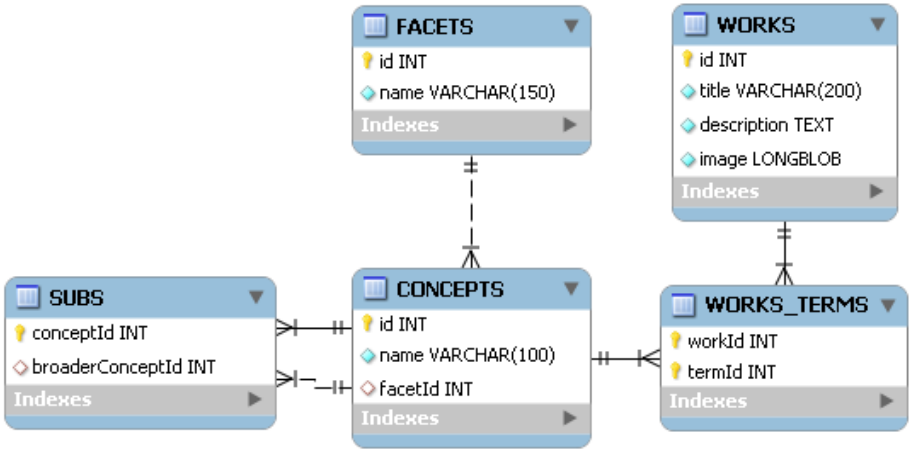

Fig. 11. EER Model for store a facet taxonomy and artworks objects**.**

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.

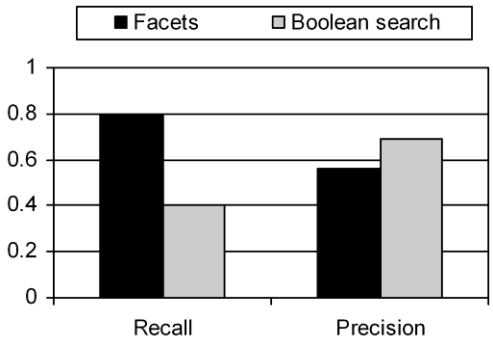

Fig. 12 Recall and Precision evaluation.

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.