An Approach to Art Collections Management and Content-based Recovery

Concepción Pérez de Celis Herrero*, Jaime Lara Álvarez**, Gustavo Cossio Aguilar* and María J. Somodevilla García*

Abstract—This study presents a comprehensive solution to the collection management, which is based on the model for Cultural Objects (CCO). The developed system manages and spreads the collections that are safeguarded in museums and galleries more easily by using IT. In particular, we present our approach for a non-structured search and recovery of the objects based on the annotation of artwork images. In this methodology, we have introduced a faceted search used as a framework for multi-classification and for exploring/browsing complex information bases in a guided, yet unconstrained way, through a visual interface.

Keywords—Search by Content, Faceted Classification, IT, Collections Management, Metadata, Information Retrieval

1. Introduction

Currently, many art museums offer descriptive catalogs as a part of their websites. The online catalogs assist in the spreading of artistic collections and facilitate their access by researchers and curators. For such purposes, multimedia databases are commonly used for the storage of images and information about artistic objects.

Today, the safeguard of cultural heritage has become a priority topic, because it constitutes the foundations of the identities of towns. The first step in the conservation of this heritage is, undoubtedly, registration, and cataloging. This may result in the protection, spreading, and appropriation of heritage by current societies and preservation for future societies.

It is mandatory to fulfill the standards in the cataloging context and all that is related to the procedures to carry out the relative issues in the design and management of expositions, which are nowadays one of the main activities in almost every cultural institution.

The need to analyze, describe, organize, and recover images is now more frequent. Nowadays, one can find several proposals for the indexing and classification of images that are mainly centered on the following extractions of objective data from images: colors, shapes, and identifiable elements such as people, objects, animals, etc. All of these make it easier for greater possibilities for recovery like: text-based recovery, content-based recovery, or a combination of both, namely, content-text-based recovery.

Historically, art curators have retrieved images by first manually annotating them with key-

Manuscript received February 11, 2011; first revision May 20, 2011; accepted June 13, 2011.

Corresponding Author: María Josefa Somodevilla García

^{*} Facultad de Ciencias de la Computación, 14 Sur y Boulevard San Claudio, Cp.72570, Puebla, México (mcpcpelish@gmail.com, gcossio@cs.buap.mx, mariasg@solarium.cs.buap.mx)

^{**} Área de Educación a Distancia, Universidad Tecnológica de Puebla (jimmy.tec@gmail.com)

words. Given a query, these annotations are used to retrieve appropriate pictures. Underlying this approach is the belief that the words associated (manually) with a picture essentially capture the semantics of the picture and any retrieval based on these keywords will, therefore, retrieve the relevant pictures. Since manual image annotation is expensive, there has been great interest in producing automatic ways to retrieve images based on content.

In this paper we will present an integrated management art collections system. We will also address the problem of art image retrieval (AIR) and introduce a methodology to reduce the semantic gap in the annotation of art images from a large art image collection.

2. CATALOGING STANDARDS

One of the main standards for the cataloging of cultural objects is the one proposed by the Getty Research Institute, namely, Categories for the Description of Works of Art (CDWA) [1]. Although it is a very complete and accepted model, it does not, from our very own point of view, comply with all of our possible descriptions of complex objects that are required in the particular case of industrial heritage objects and new media art.

A second model considered was the Cataloging Cultural Objects model (CCO). A guide to describing cultural works and their images was published as a draft in February 2005, and its final version was published in 2006 [2]. This model, is not specialized in the works of art, but is applicable to any type of cultural object, thus, making it a model that can be applied to describe complex objects.

A relevant feature of the cataloging model proposed by CCO is the possibility of deriving a metadata model from it. This metadata model enables the carrying out the description of the object to be cataloged in a detailed way, depending on the needs of the cataloger. For our particular case, this metadata model allowed us to perform the necessary extensions to present a proper description of the so-called complex objects.

3. METADATA

When referring to metadata, we mean data about the data. The global interpretation of this concept allows us to visualize all that can be said about one object that can be manipulated as a discrete entity. In this way, the object in question can be manipulated as a single entity or as an entity that is formed by the grouping of several objects, which can be simultaneously considered as single entities. Generally, irrespective of whether it is a physical or conceptual object, we can refer to it through metadata in the three following aspects [3]: Content, Context, or Structure.

- Content: refers to the content of the object, which is what causes it to remain as what is intrinsic to the object.
- Context: indicates when, how, why, and under what circumstances an object can be created or eliminated. This is intrinsic to the object.
- Structure: establishes the associations of the object to other objects. These associations can be conceptual or physical and can be intrinsic or extrinsic to the object.

The metadata can thus be complex constructions that need to be created and kept, and that re-

fer to a specific group of objects. The CCO model particularly focuses on the aspects of the content and context previously described, to characterize the objects related to cultural objects. The CCO covers several types of cultural objects, which includes architecture, archeological sites, and devices, but its fundamental emphasis, even though it is not exclusive, is on artistic objects: paintings, sculptures, printings, manuscripts, photographs, and other visual means. It is especially designed to describe the different collections that belong to art and architecture museums.

To use the model proposed by CCO in the collections of industrial matters, some modifications and extensions need to be done. There are two types of visualized extensions. The first one is related to the cataloging attributes of the objects (this falls directly into the metadata model). For this, concepts such as *heritage type* and attributes, like *use, maker, and serial number*, are added.

The other groups of extensions considered are related to objects modeling. For this modeling, we will use the complex object to model the user's perception about the entities of the real world. Through this way, we define the following:

- Simple Conceptual Objects: They are objects that are visualized as a single object, even though they are not the only objects. Well-defined objects can be identified with the following types:
- 2. Item: a simple object. It can be that other objects, which are called components, form it. Such components must be cataloged. The items formed by components are added objects and can be created as a result of the operation of applied addition on already existing objects. The existing relationship in these types of objects is a *component-of*.
- 3. Set: refers to a concept of the real world that associates a group of possibly heterogeneous objects. Such objects are not found individually cataloged, but in a group. A group, according to CCO, is an addition of items that share the origin. A group can be formed by hundreds or a few items (e.g., the drawings, models, and documents in an architect's office).
- 4. Series: a complex object that is formed by the association of items, that are not necessarily homogeneous, whose association can be arbitrary. In other words, it is an extrinsic premise. These days, a series in CCO is a number of works that were created with a temporal succession by the same artist. It can also be a study that has been set aside for being exhibited, according to the creator(s), together or in a succession in a cycle of works. The works of a series typically shares related topics, the same means, or similar or any other type of characteristic, and they are conceptually related. Every work within a series must be in separate records and as a whole. The relationship between the work and its series must be intrinsic, because in this way, the work is better understood within the context of the series.

4. TESEO: Information System for Cultural Objects

The idea to develop a system that can allow the management of cultural heritage collections has been explored in the last years. This is the reason why some versions of this kind of system already exist, for example, the Spanish DOMUS [4] or the French JOCONDA [5].

The integrated management of collections system, TESEO, subsequently arises as a response to the real need for cataloguing and managing the collections that belonging to different cultural

institutions in Mexico that require a tool, such as, the first step towards the protection, conservation, and dissemination of the country's cultural heritage.

In its origins, the project aims to meet the standards in the areas of documentation and administrative process (incoming items, borrowed items, and internal moves) and also in all the activities related to the procedures to the design and management of expositions, which are currently one of the main activities inside almost every cultural institution. For this reason, it is is useful to develop a tool that assists in the activity of cataloguing, and to reflect the processes of the curators, exhibition coordinator, museographers, conservators, and restorers.

The TESEO project has the following subsystems: Cataloging, Administration, Search Engines and Exposition Design. We only detail the modules of the Cataloging and Search engines since both modules are related to the context of this work

Table 1. Example of cataloguing with a TESEO system, the original text is in Spanish, the translation was done for explanation motives

METADATA AND TEXTUAL DESCRIPTION

<work>

<recordType>item</recordType>

<workType>Pintura</workType>

<title>

<display>Autorretrato con modelo</display>

<titleType>Preferido</titleType>

<language>Español</language>

</title>

<creator>

<contributor>

<name type="real">Christian Schad</name>

<controlledCreatorsource="AAT" target="500030368">

<date type="life">

<earliest>1894</earliest>

<lastest>1982</latest>

</date>

<role>Pintor</role>

</contributor>

</creator>

<meassurements>

<display>76 x 61,5 cm</display>

<data value="61,5" unit="cm" type="ancho">

<data value="76" unit="cm" type="alto">

</meassurements>

<materialTech>

<technique>Óleo sobre madera</technique>

<materials>

<material targetID="42568">Óleo</material>

<material targetID="01256">Madera</material>

</materials>

</materialTech>

<culture>Alemana</culture>

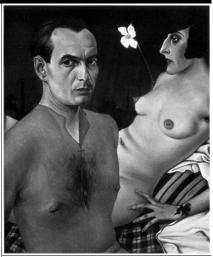
<date>

<earliest>1927</earliest>

<lastest></lastest>

</date>

</work>



The artist's transparent shirt reveals his chest. He is positioned in front of the woman, but only partially conceals her nakedness. A diaphanous curtain separates them from the city. Schad's precise realism is loaded with symbolism. A narcissus, indicating vanity, leans towards the artist. The woman's face is scarred with a *freggio*, inflicted on Neapolitan women by their lovers to make them unattractive to others. It is a startling emblem of the potential violence underlying the male possession of the female body.

- a) Cataloging: for this module, the cataloging mode proposed by the CCO group was considered. Some extensions to the model proposed by this group were implemented. In particular, the possibility of modeling complex objects, that is, objects that are simultaneously formed by other objects under the addition operation for which they are a part-of. In this way, it is possible to carry out the proper description of groups (as in the case of a description of a system), series (e.g., a collection of paintings related by a specific characteristic by the artist), or to simply describe the fact of serialization in the edition of engravings. In the description of the pieces, everything related to their history and physical condition are included, as well as indicators to the expositions in which they have participated.
- b) Search engines: some search engines have been developed depending on the different user levels. These search engines allow for the retrieval of information not only by key words, but also by the concatenation of characteristics that allow the relation of advanced searches. Nowadays, some works are being carried out to establish an interface that allows us to do a search by concepts about the content textual descriptor of the piece. This aspect, which is the aim of this research, will be discussed later in the following sections:

Table 1 shows a cultural object and its textual description. This object will be used throughout the methodology to exemplify the steps for classification.

5. Information Retrieval of Artistic Objects

In the description of the four works of art presented in Table 2, information retrieval could be done by the cataloging descriptors: author, title, or pictorial tendency. One of the objectives that TESEO tries to achieve is to provide, besides the metadata search engines, retrieval image engines with different kinds of features like: Denotation, Connotation or Genre.

In other words, if a user was interested in retrieving pieces that include women, three of the pieces should come up as well as all of the pieces that contain female figures. On the other hand if the user searched for pieces that express *movement*, the "Dancer" piece of art should be in the answer set as well.

The type of search pointed out, implies the use of facets that allows for the proper retrieval of items searched. The types of query required should then allow for the retrieval of information, either through the already-existing facets in the system or should be defined at the time of execution through search equations in which the user is able to include a series of words that may indicate the elements that define a concept, which must be found in the description of the piece. For example:

Urban = vehicles + buildings + people

The defined concept of **Urban** may become permanent upon the user's request. By doing so, some feedback can be obtained from the user. On the other hand, as an object may be attached to more than one concept, it would be able to categorize the order of the importance of the object regarding the concepts related to it.

To support the structure and efficient retrieval of the artistic objects, and given the complexity in the multiplicity of the possible interpretations and contents that can be linked to the objects,

Table 2. Artist object description, the original text is in Spanish, the translation was done for explanation motives

AN EXAMPLE OF ARTISTIC OBJECTS 2 Larry Rivers. Parts of the body: Pablo Gargallo. French vocabulary Dancer, lesson, 1934. 1961 - 1962. Welded copper Oil painting on Height: 78 cm. canvas 183 x 122 cm. Comment: A dancer performs a pirouette and it Comment: Several parts of a body and the naked causes her skirt to fly. The welded copper shape shape of a female. It was painted schematically. The gives an impression of lightness, freedom, and parts of the body have been handwritten, as if it were dynamism. The artist has omitted all peripheral a French vocabulary lesson. This work of art stands details, and has focused on capturing the essence of as an erotic portrait and pedagogical illustration. the movement of the dancer. Dorothea Tan-Ben Vautier. ning. L'ART EST INUTILE Art is useless, let's Music at night, go back home. 1946. 1971 Oil painting on RENTREZ CHEZ VOUS Acrylic on canvas canvas. 97 x 130 cm. 41 x 61 cm. Comment: A girl in a white shredded dress, hair flying upwards, is in front of a sunflower, whose << Art is useless, let's go back home >>, reads the broken stalk moves towards her, like a giant green text in French. The attractive text is printed in simple tentacle. In this kind of mournful and untidy hall, white letters on a scarlet red background. the girls look they are static and petrified in the time

we recommend the use of a faceted classification process.

Under the previous premise, faceted taxonomies can take advantage of the ways in which the metadata act. Metadata are the collection of structured information about any type of object or any part of that object, (e.g., the name or title of a work art, the author, age, image, textual description [denotation], its interpretation [connotation], or its gender). All of these are some kind of metadata associated with a cultural object.

In our approach, we have implemented a faceted classification, and thus, the metadata model proposed was extended with the metadata of its genre, connotation, and denotation, so as to include its textual description as a part of the requests of the objects. This fact determines the need for a methodology that allows us to generate the taxonomic structure of the facets on which the different terms included in the textual description of the objects are going to be classified.

In this natural way, every element of the metadata structures can be incorporated as a concept of the faceted taxonomy and can be retrieved through a search engine. The importance of tax-

onomies is based on the possibility of their use as a triplet (classification diagram, semantic interpretation, knowledge map) [6] in the organization and later retrieval of possibly stored objects in a database. As a consequence, we have the opportunity to access the request of an object under any of the dimensions in which they were classified. If we consider the example previously given, we could retrieve a cultural object by its gender, connotation, denotation, or simply surf for the different facets, because they are orthogonal among themselves.

Sacco [7] introduced the concept of dynamic taxonomy, along with the notion to bear the incorporation of facets that they themselves require an independent taxonomy description. *Dynamic taxonomies* [8] (*DT*, also recently known as *faceted search systems*) are a general knowledge management model that is based on a multidimensional classification of heterogeneous data *objects*, and are used to explore/browse complex information bases on a guided yet unconstrained way through a visual interface. The model is primarily concerned with user-centered access, and object classification is not addressed in the base model.

As our interest is the retrieval of the works of art based on their content we used the textual description of the images for their annotation. We will use the fundamentals of the dynamic taxonomies, but will also extend the domain of data modeling established by Sacco, including textual objects within the metadata.

In literature, we can find two interesting approaches for the treatment of textual objects. The first one presupposes the existence of several topics on which taxonomy is to be developed, and in this way, the job of the algorithms developed is to extract the terms and concepts linked to the topics in question from the documents. In this approach, each of the several taxonomies derived requires a group of key terms that are linked to the concepts based on which the classification of the documents of the target objective will be carried out.

The second approach, in contrast, focuses on determining the facets and defining their taxonomy, which is a departure from the textual analysis of the documents of the collection. In this case, some techniques for text analysis are used. In particular, in [9], is described an algorithm, used by the Flemish project at the University of Berkeley for the automatic generation of facets about the corpus in the English language, using WorldNet.

6. FACETED INDEXING AND CLASSIFICATION

The purpose of indexing is to identify the concepts that represent the content and translation of documents into a way that can be computably managed [10]. Use of a faceted classification will assist us in analyzing the text, and will allow us to ensure that we are entering all relevant items that should be indexed.

We will now present a brief description of the three facets of genre, denotation, and connotation, which we already mentioned in the previous paragraphs. Using these three facets, we would then extend the classification of the artistic objects.

- Genre (artistic subject): each of the different categories or kinds that can be used to order
 the works depending on common features, such as shape and content: portrait and selfportrait, landscapes, religion, mythology, etc.
- Denotation (content): concepts that are shown in an artistic work.
- Connotation (interpretation): concepts reflected or transmitted by the artistic work. For ex-

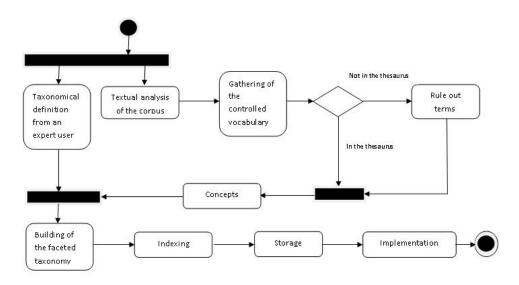


Fig. 1. Activity diagram of the dynamic taxonomy generation process

ample the word rain has the connotative meaning of sadness or melancholy.

From these three facets, only an expert user defined the facet genre. The other two facets of denotation and connotation were obtained automatically.

In order to develop an image annotation system and a motor search for the retrieved objects in a collection of 500 works of art using their textual description in Spanish [11], our methodology is based on some techniques and algorithms used in Information Recovery (IR) to generate in a semi automatic way a faceted classification that allows us to link the resort of a document to the different facets. By doing so, we could obtain a controlled vocabulary extracted from the processing of different texts of the target collection. Subsequently, with the help of a controlled vocabulary and thesaurus, the concepts were determined. These concepts were structured hierarchically with the aid of a thesaurus. In particular, we used the online thesaurus of human emotions called the *Human Emotions Thesaurus* (HET), which was developed by clinical researchers of the University of British Columbia and the *Tesauro Europeo de la Educación*, structured in Spanish and created in 2003 by the Commission of the European Communities, which contains 3,516 terms and 2,100 relations.

Finally, the objects are indexed [12] with the faceted taxonomy. This allowed us to generate a faceted classification system. The objects, the faceted taxonomy, and indexes are stored in a database and integrated into a system that enables the user to explore and, at the same time, refine the search through a navigation tree. In Fig. 1, the process is shown through an activity diagram.

7. USER INTERFACE DESIGN

Exploration by the user will be realized by means of a navigation tree [13] (taxonomic tree),

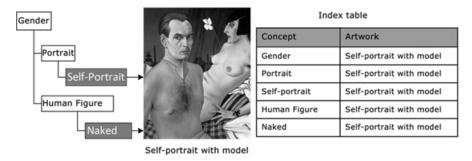


Fig. 2. An object with multifaceted classification and the index table

which guides the user in his/her query. In addition, it allows for the structuring, presenting, and acceding of the information in more than one dimension. This benefits the user, because the user can easily and intuitively locate and explore the information by means of the different approaches provided by the facets. In Fig. 2, we have shown an example of an object classified under two facets.

One of the core principles of dynamic taxonomies is to restrict the available filtering options in the given focus to only those that will lead to a nonempty result set. Hence, the user can never run into a situation with zero results. This is opposed to the process in a typical advanced search situation, where a complex boolean query is first constructed, which is then evaluated by demand. This, however, can result in empty result sets, often without further indication on which part of the query could be relaxed to retrieve some results. The exclusion of potentially frustrating situations by design is often referred to as the "Poka-Yoke Principle" (see Fig. 3).

In Fig. 4, we summarize our approach for automatic image annotation of an art collection of pieces that is based on their textual description. In this way, every piece of art should be indexed not only with the information contained in textual descriptions, but should also have the hierarchical information of the concepts that lie in the dynamic taxonomy.

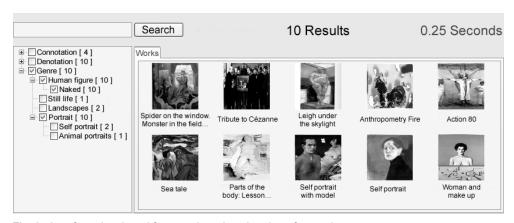


Fig. 3. Interface developed for search and exploration of artworks

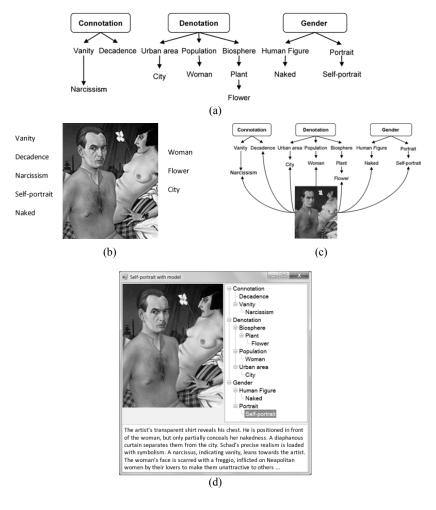


Fig. 4. Summary of the approach: a) an example of the facets with their taxonomies b) terms in the controlled vocabulary to be indexed to the object c) taxonomy-related concepts and d) indexed art work and view in the user interface

8. RESULTS AND FUTURE WORK

In this paper we have presented a system for managing and cataloging cultural interest art objects. In particular, an idea was also developed for the automatic image labeling of collection pieces that is based on their textual description has been included as a part of their cataloging metadata.

The methodology we propose allows for multi-classification. Therefore dynamic taxonomies were used and three different facets were built: Gender, Denotation, and Connotation.

We made and implemented algorithms that allow for the retrieval of images by relying on their contents and thereby contributing to solving the semantic labeling of images, which is one of the goals of the Content-Based Image Retrieval methodology (CBIR). The concept extraction process and indexing between the facets and the works was designed and subsequently implemented in the AWK text-processing programming language. This module can be used on any set of textual descriptors

With regard to the construction of dynamic taxonomies and navigation, the following operations were implemented: Zoom-In Zoom-Out, Pivot, Shift, and Query by Example (QBE).

Dynamic taxonomies that were generated, unlike those reported in the literature discussed, are not on a closed domain. An example is the Denotation taxonomy, in which there is an open domain that is associated with a thesaurus for expanding the domain of the facets obtained from the analysis of unstructured text.

Among the extensions and improvements that we have referred to, are those of incorporating the following: "The Art & Architecture Thesaurus" ® developed by the Getty Research Institute (GRI), and translated into Spanish by the Documentation Center of Patrimonial Assets (under the Directorate of Libraries, Archives, and Museums); the "Thesaurus Spanish Historical Heritage"; the "Multilingual Thesaurus of Nations," and the "Extended Multilingual Wordnet" (University Alicante).

REFERENCES

- [1] J. Paul Getty Trust, College Art Association, Categories for the Description of Works of Art (CDWA), http://www.getty.edu/research/conducting_research/standards/cdwa/index.html
- [2] Baca, M.; Harpring, P.; Lanzi, E.; Mcrae, L.; Whiteside, A., Eds. Cataloging Cultural Objects: "A Guide to Describing Cultural Works and Their Images"; American Library Association, Chicago, 2006.
- [3] Gilliland-Swetland, Anne J. "Setting the Stage: Defining Metadata" in Introduction to Metadata: Pathways to Digital Information, Murtha Baca, ed. (Los Angeles: Getty Information Institute, 2000), second edition. 12 pp. (Third edition summer 2006).
- [4] DOMUS, http://www.mcu.es/museos/CE/Funciones/Documentacion/DocumenatacionColecciones.html
- [5] "JOCONDE" Catalogue des collections des musees de France, 2004.
- [6] Lambe Patrick, "Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness". Chandos Publishing, 2007.
- [7] Sacco, G. M., "Dynamic Taxonomies: A Model for Large Information Bases". *IEEE Transactions on Knowledge and Data Engineering* 12:2, 2000, pp.468-479.
- [8] Sacco, Giovanni Maria; Tzitzikas, Yannis, "Dynamic Taxonomies and Faceted Search Theory, Practice, and Experience"; The Information Retrieval Series, Vol. 25 Sacco, Giovanni Maria; Tzitzikas, Yannis (Eds.) 2009, XVII, 340 p., Hardcover ISBN: 978-3-642-02358-3.
- [9] Stoica Emilia, Hearst Marti Y Richardson Megan. "Automating Creation of Hierarchical Faceted Metadata Structures". In the proceedings of the Human Language Technology Conference (NAACL-HLT), Rochester NY, April, 2007, pp.244-251.
- [10] Tolosa Gabriel Y Bordignon Fernando. Introducción a la Recuperación de Información Conceptos, modelos y algoritmos básicos. *Universidad Nacional de Luján, Argentina*. Available at: http://www.tyr.unlu.edu.ar/tallerIR/2008/docs/Introduccion-RI-v9f.pdf.
- [11] "El ABC del Arte Del Siglo XX", Primera Edición En Español 1999, Editorial Phaidon Press Limited.
- [12] Leise Fred. Using Faceted Classification to Assist Indexing. Contextual Analysis, 2003. http://www.contextualanalysis.com/pub_usingfacets.php
- [13] Tzitzikas Yannis, Analyt Anastasia, Spyratos Nicolas Y Constantopoulo Panos. "An Algebra for Specifying Valid Compound Terms in Faceted Taxonomies", *Journal Data & Knowledge Engineer*ing Volume 62 Issue 1, July, 2007. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands Data, pp.1-40.



Concepción Pérez de Celis Herrero

She obtained her PhD in Computer science from the University Pierre et Marie Curie of Paris, France and is a specialist in the Management of Cultural Heritages. She has been curator of numerous exhibitions and was the Art and Producing Director of the Node of Art and New Technologies at Plataforma 06. She is currently a Professor-Researcher of the Faculty of Computer Science, at the Benemérita Universidad Autónoma de Puebla, México.



Jaime Lara Álvarez

He has his Master's in Computer Science from the BUAP (Meritorious Autonomous University of Puebla), Mex 2009. His research interests include: datamining, database systems, information retrieval, and object oriented programming. He currently is a professor in the E-learning Department at the UTP (Technological University of Puebla).



Gustavo Cossio Aguilar

He studied Computer Science Engineering and is a Developer and Webmaster and

Senior Programmer for the TESEO project. Since 2007, he has worked as a Network Administrator for the Computer Science Faculty of the BUAP.

His research interests are network security, database systems, image processing, knowledge retrieval, and computer forensics.



María J. Somodevilla-García

She was born in New York, USA, in 1958. She obtained her BS. in 1982 from the Superior Polytechnic Institute of La Habana, Cuba, her Master's degree in Deductive Databases in 1995 from the Engineering Faculty of the "Americanum University of Puebla, Mexico," and her Ph.D. degree in Spatial Databases in 2003 from the Engineering School of Tulane University, New Orleans, LA, USA. Since 1986, she has worked for the Faculty of the Autonomous Benemerita University of Puebla, Mexico. She is author and coauthor of about 40 publications on

data representation and analysis.