Catalina Naumis Peña
Researcher at the Centro Universitario de
Investigaciones Bibliotecológicas/
University Library Research Center.
Universidad Nacional Autónoma de Mexico/
National Autonomous University of Mexico

Evaluation of Educational Thesauri

Abstract: For years, Mexico has had a distance learning system backed by television-signal-transmitted videos. The change to digital and computer transmission demands organizing the information system and its subject contents through a thesaurus. To prepare the thesaurus, an evaluation of existing thesauri and standards for data exchange was carried out, aimed at retrieving subject contents and scheduling broadcasting. Methodology for evaluating thesauri was proposed, compared with a virtual educational platform and a basic structure for setting up the information system was recommended.

1. Introduction

The purpose of this paper is to evaluate educational thesauri in order to decide whether to construct a new one or adapt one that already exists. The area of expertise that will be analyzed is that of educational contents, given the need for documentary organization demanded by the current distance learning proposal. To represent educational contents for distance learning programs using multimedia digital support in information systems, it is necessary to have a documentary vocabulary that establishes the indexing terms of the educational system to be supported.

Distance learning systems require secure, precise and efficient shared information. While the problem at hand is the evaluation of thesauri, we should start with some background on information retrieval mechanisms in the educational contents of the Web.

2. Use of Automatic Computing for Educational Content Information Retrieval

Distance learning relies on computer intervention to manipulate the information that each particular educational program needs. E-learning systems combine educational resources or learning objects, with people’s pedagogical activities, in their respective roles (Griffiths, Blat, García, Sayago, 2005, 6). Our concern in this paper are the learning objects related to educational subject contents, their indexing and retrieval. In the automatic computing arena, contents plus metadata are the essence (Jong, 2003, 9), so it is essential to clarify that the contents we observe here are educational, not automatic computing, subject contents.

Metadata are used to locate, identify, select and have access to the learning objects. Metadata also document how that object behaves, its function and use, relationships with other objects and visibility characteristics. Data exchange standards help define the interfaces so that different metadata schemes may be transferred in educational information systems.

Whatever technology is used, it must enable the information exchange to retain the metadata’s meaning and original structure. These standards offer a neutral representation of the metadata and their structural order. They have nothing to do with the underlying semantics but rather with providing a common, machine legible way to transfer the elements that have been defined through networks, systems and platforms. Data exchange standards thus provide the mapping interfaces between the level of definition and the technological level of information systems (Jong, 2003, 16)
In general, educational information systems are using the Extensible Mark-up Language (XML), which distinguishes between form and content (unlike HTML), to translate any metadata scheme to a common representation format, with the aim of transferring them through the Web. The Resource Description Framework (RDF) works like a global metadata exchange structure and provides the base work for other standards. RDF has defined a top level metadata model and syntax that is expressed in XML. The RDF model presents three object types:

- a resource; anything that can have a URL address
- a property: a resource with a name that can be used as a property (e.g. author, title)
- a statement: the relationship between a resource and a property

While objects are handled in the RDF model, the relationships among them are not the ones needed to develop the full potentiality of the query systems and arrive at the learning objects. Other specifications, also based on object programming, have been proposed, in which the information search explores within content. Outstanding among the proposals are the IEEE (Institute of Electrical and Electronics Engineers) Learning Technology Standards Committee Learning Objects Metadata Working Group, the Dublin Core Metadata Initiative, the Global Learning Consortium which proposed the IMS Resource Metadata and the Advance Distributed Learning/Sharable Content Object Reference Model. Some of the people involved in the development of these standards comment, “to develop mutual interoperable metadata for technology-supported learning, education and training tools” (Hodgins, 2001, 1), “...having a common approach to educational metadata is crucial to further speed up adoption of metadata technologies. That, in turn, is the first, crucial step on the long road to open learning infrastructures” (Duval, 2001, 2)

The virtual platforms mentioned evolve, and for example Dublin Core, which started out providing bibliographic metadata for information resources and dealing with text resources, has introduced certain extensions to describe multimedia and audiovisual resources that cover specific audio and video aspects.

The metadata model should logically connect metadata entities, relationships and attributes to the essence (digital) itself (Jong, 2003, 9). The model will have to link the information elements to the description of a whole that encompasses them and define the level of access within the contents. Another feature of the standard is that it be able to link textual, and sound information, as well as fixed and moving images.

3. Representation of Educational Subject Contents

As an offshoot of the previous point, educational subject contents in the realm of automatic computing are learning objects organized into virtual educational platforms. In addition to the interface model for perusing and transmitting a general idea about the location and characteristics of the subject contents, is the problem of linguistically representing those contents.

Linguistic representation of educational contents is necessary both for textual digital documents and multimedia. The wealth of expressions makes the former confusing when they are consulted in a concept rather than word search. So as not to repeat an expression in a single educational text, words that are not habitually used by the educational platform users may be used. The process becomes long and complex when the user does not find the theme and must search all the possible forms in which the same concept may have been expressed ...

“When searching for information about concepts that can be expressed in multiple ways it is more effective to use classified Web directories. A directory controls for synonyms and
homonyms and provides context for the index terms by placing them in a hierarchical structure.” (Mai, 2004, 92)

Thesaurus descriptors are another kind of specific content metadata and are used mainly as an indexing tool. In digital multimedia documents, linguistic representation is basic for locating images that will illustrate an educational content consisting of zeros and ones. Educational images have to transmit exactly the subject meant to be taught so as not to confuse the student, and although it has been said that, “A picture is worth a thousand words,” the collective experience of information professionals assures that a word can take us to an image, yet the opposite does not always happen.

In today’s technological world, many access points may be used ...“the computer is capable of storing and organizing texts in ways that enhance their retrievability, and the ways that knowledge can be organized in digital archives on the Internet by using a variety of potential access points” (Andersen, 2002, 37). This fact, which Andersen explains, is a great benefit, becomes a problem when too much information is retrieved. Discretion is not an exclusively human virtue; it is also desirable in information systems.

The virtual teaching platform does not solve conceptual information retrieval if words in the text are used as the sole guide. The support of a thesaurus is basic for specifying themes consistently. However, the creation of a thesaurus is costly and difficult, so that proposals available on the market should always be analyzed as possible knowledge organization solutions.

4. Methodology for Evaluating Thesauri

In Mexico, the need developed for a thesaurus of educational contents with a classification structure, in order to retrieve videos for use in the transmission of distance learning television programming at the elementary and intermediate levels. The first task was to analyze existing educational thesauri, and a methodology was set up that would be applied to each one in the same manner.

This evaluation is carried out within a particular field and type of information, so as to zero in on a practical problem, for which the consistency and balance among existing thesauri is evaluated alongside the proposal of a new thesaurus. Given the characteristics of this paper, it is not possible to delve into details, but the methodology applied and the result of the evaluation are summarized. The proposed methodology focuses on five different points of view: presentation analysis, consultation analysis, consistency, content analysis and terminological and semantic structure.

All five of these perspectives are important: presentation, because while the main part of a thesaurus contains an alphabetical body of descriptors and indices with entries for indexing and retrieving information that differ from the alphabetical body, there must be an introductory explanation of sorts regarding a series of aspects to initially help users decide whether it is useful for the information system meant to be inserted and later take full advantage of the tool being offered. It has to do not only with greater or lesser ease of handling but with the authors’ explanation as to the level of conceptual structuring of the thematic field of application.

Consultation defines the way of accessing the thesaurus, whether on line, on a compact disc or in printed form, as well as the date the terms included were updated.

Consistency involves observation of the following elements: reciprocal relationships, terms in the same form and under the same circumstances, use of clarifications on the application of terms, standardization of the genre and translation to another language.

Content deals with the relationship between the number of descriptors and non descriptors, the levels of hierarchy and the types of contents included.
The terminological and semantic structure of the thesaurus entails the analysis of the thesaurus’ subject division, in other words, terminology dissection and organization.

Applying a methodology for reviewing existing thesauri is not only done to discard the possibility of using a preexisting one instead of creating a new one but with a further intention of reconciling partially problematic aspects of the thesaurus to be generated, in addition to comparing the criteria used in the design of the most general categories and, of course, take advantage of their useful elements for the thesaurus to be developed.

So, while "a thesaurus for organizing educational contents must be the faithful reflection of the educational system which it is going to serve" (Naumis, 2002, 9), the following educational thesauri were analyzed, for the methodological and practical advantages already explained:

- Thesaurus of the UNESCO: structured list of descriptors for bibliographic indexing and retrieval in the spheres of education, science, the social and human sciences, culture, communication and information / United Nations Education, Science and Culture Organization. – http://databases.unesco.org/thessp/

Selection of these thesauri was mainly based on their general focus (in other words, they deal with education), the concrete aspects undertaken and their present use (they are recent thesauri that include audiovisual material and are centered on the use of new didactic resources).

The framework used for the analysis of the thesauri was prepared in accordance with certain indicators which were taken, adapted and complemented among themselves, in keeping with works by Alvaro Bermejo(1989) and Michel Dauzats (1994), and with our own indicators added.

The review of thesauri from which to select the sampling made it evident that most have been produced by international institutions, with fewer thesauri having a national radius of action, and are limited to the scientific realm, not even dedicated to supporting a teaching system.

The field of the thesauri analyzed may be summarized as follows: economic and social development, education in an international context, science and technology for development, United Nations Programs, another one with various areas (education, science, culture, social and human sciences, culture, information and communication, politics, laws and economics, countries and human groups) and, finally educational research.
The UNESCO thesaurus, put together by the International Office of Education, is divided into eight semantic fields, and in principle the faceted organization should be centered on educational themes. However, upon examining the contents of the semantic fields, it was found that they actually encompass UNESCO projects and activities. Even the third section, devoted specifically to teaching-related terms, that is, knowledge transmitted through a teaching process, includes overly general terms for an educational information system. Unfortunately, the next UNESCO Thesaurus (divided into microthesauri) has the same particularities and the same limitation as the previous one.

In the UNBIS Thesaurus, which has a mainly thematic structure, not all the descriptors are part of the hierarchical order, so that it does not comply with the interlinked structure required to give the thesaurus greater consistency.

The SPINES Thesaurus, on the other hand, is organized into 34 subject fields, with all the descriptors belonging to one of them. Consequently, it is a thesaurus of educational contents in limited realms. The educational level it represents is professional and up, and the limitation observed for its functioning in an elementary and intermediate educational system level is the scientific language and subject coverage. In conclusion, it may be adopted as a model for specific educational subject realms using suitable language for user level.

The OCDE Thesaurus is an excellent model for economic- and social development-related aspects, since it is structured into nineteen classes and seven hierarchies.

ERIC is, unquestionably, the most well-known thesaurus on educational internationally, having become the international information system’s backbone on education. Nevertheless, a quick look at its content reveals that it is aimed at educational research, not educational contents. Furthermore, this linguistic tool is written in English, so it would have to be translated into Spanish before it could be used. As Grijelmo says, “Language constitutes the most faithful core of every community, and therefore no other language may define us” (Grijelmo, 2002, 283). In the same vein, Yule states that, …“your language will give you a ready-made system of categorizing what you perceive, and as a consequence, you will be led to perceive the world around you only in those categories.”(Yule, 1998, 280).

The conclusion reached through the analysis of the aforementioned educational thesauri is that none really responds to the indexing needs of a system with multimedia support documents for an elementary and intermediate level educational program.

Other solutions were subsequently analyzed so as to propose the classification scheme of a thesaurus with educational contents, and a technological proposal was found, “ROSA: Repository of Objects with Semantic Access for e-Learning”, which is based on the organization of the educational system itself. The first classification scheme is grounded in the programs and different teaching levels; in the next phase, the names of the different courses are included, and the subject classes are the contents of each one of the courses. While the information organization uses the same institutional premise, it was not possible to review and evaluate the contents since the system does not appear on line.

5. Conclusions

The terminological and semantic evaluation of the theme of the thesaurus is the most significant aspect; the others contribute to its appreciation. Beyond the result of the evaluation carried out, it is necessary to recognize that the methodology enabled approaching and understanding thesaurus organization and provided elements of help for creating a thesaurus. It can practically be ensured, however, that it is difficult to find a thesaurus with a structure of useful knowledge for any other case. Every educational space has its objectives, level of delving, vocabulary used, language and organizational traits of the institution to which it pertains.
In the analysis of library science literature on the subject and starting from the premise that the thesaurus is a classification scheme, elements were found that support the conclusion of the difficulty of reusing a thesaurus developed out of one context in another:

A classification scheme is just one potential way to describe a particular domain or the universe of knowledge. To create a classification system for a particular company, organization, library, or any other information center, one needs to begin with a study of the discourse and the activities that take place in the organization or domain. One needs to learn the language used in the community, since the classification must reflect and respond to this particular discourse community. A classification is not something that can be created for an organization by an epistemic authority; a classification must grow out of the organization. The classification is a typification of the language in the organization. (Mai, 2004, 46)

The difficulty lies in the very social, cultural and economic dynamics of the institutions where the need arises for organizing the contents they generate ...“concept of knowledge organization is in interaction with and derived from the social organization of knowledge.” (Andersen, 2002, 37)

In view of the previous conclusions compared with elementary and intermediate education in Mexico we propose organizing the classification scheme of the thesaurus for multimedia educational contents with different types of metadata: the first ones to reflect the different educational levels taught, others for the different courses encompassed by each educational level and the last ones for the educational contents, as such, that will be developed through the usual thematic interrelationship in a thesaurus. The technological part must take care of the relationships among the three types of metadata, to enable the option of retrieving any educational content, in order to open the option of using them in cultural promotion or extension programs.

Reference List

Repository of Objects with Semantic Access for e-Learning

Le thesaurus de l'image : etude des langages documentaires pour l'audiovisuel / Sous la direction de Michel Dauzats. París: ADBS Editions