Computer Science and Information Technology 1(1): 19-32, 2013 http://www.hrpub.org
DOI: 10.13189/csit.2013.010103

Conceptual Database Retrieval through Multilingual Thesauri

E. Petraki, C. Kapetis, E. J. Yannakoudakis*

Athens University of Economics & Business Department of Informatics, Athens 10434, Greece
*Corresponding Author: eyan@aueb.gr

Abstract In traditional database management systems, information retrieval is often carried out using keywords contained within fields of each record. Because a term (concept) can be expressed in several ways, a significant number of records are ignored by the free text techniques which use only *a posteriori* relations between terms. This paper proposes the utilisation of *a priori* conceptual relations between terms that exist independently of any documents through a controlled vocabulary known as *thesaurus*, which incorporates both terms and the conceptual relations among them. The paper discusses the integration of multilingual thesauri in the set-theoretic FDB (Frame DataBase) data model, which offers by default a universal schema for all applications. All changes to the structure of the logical-level database schema can be carried out by modifying the appropriate metadata. The purpose of this extension is for the database user to be able to apply queries on a database using information through multilingual thesauri. This approach extends the FDB model so that users can apply queries to the database using both *a priori* and *a posteriori* relationships. Apart from free text retrieval and "conceptual searching" the proposed structure enables multilingual searching independently of the language used to store data itself.

Keywords Multilingual Thesaurus, Databases, Information Retrieval, Universal Schema, Conceptual Retrieval

1. Introduction

In a traditional RDBMS users apply queries to the database via query languages like SQL. This technique requires users to have knowledge of the database schema and use a query language to search information; this search model is complicated for most ordinary users [10]. Queries using keywords is the most widely used form of querying today while it is used to search documents on the Web [11]. Keyword query is easy and flexible because it doesn't require from the database user to know details about the underlying schema. On the other hand, search techniques through keywords use ranking mechanisms in order to rank more or less relevant (to the keywords) answers. This type of utility is missing from most database management systems today while all the tuples retrieved have the same significance. There is a number of research and commercial systems that support keyword search and browsing in relational and semi-structured databases, such as DataSpot, EasyAsk, DISCOVER, BAKNS etc [11], [12].

The goal of information retrieval is to identify documents which best match user needs [8]. In Database Management Systems, information retrieval is often based on free text techniques which operate on string matching and ignore any conceptual information. This is in effect a technique for the identification of records based on a set of words contained within each record. Since a term (concept) can be expressed in several ways, a significant number of records are ignored by the free text indexes which use only *a posteriori* relations between terms. Query languages such as SQL (Structured Query Language) and its extensions are designed to query data contained in databases using pattern matching and free text techniques while conceptual information is ignored. Evidently, we need more efficient and intelligent retrieval systems which expand user queries automatically with related words. We aim at designing a system which automatically discovers related terms to expand user queries. Terms come from three main sources: 1) query specific, 2) corpus specific and 3) language specific and provide a richer representation of the user's query [8]. Alternatively, we can accommodate and utilize a knowledge base between the user and the database in order to enrich queries with more related words, thus offering conceptual retrieval rather than pure text matching [9].

The purpose of our work is to exploit the information provided by a multilingual thesaurus in order to expand user queries applied to a database with relevant terms derived from the thesaurus. The adoption of a thesaurus provides the means to express *a priori* (semantic) relationships in order to document information generally. Extending free text searches using multilingual thesaurus relations, the set of retrieved records grows as a result of similar terms expressed in several languages.

Among the terms used for documentation and retrieval

information, there are two basic categories of relations, namely the *a priori* and the *a posteriori* [1]. The *a posteriori* relations among terms are used to identify the subject of a document. For example, for an essay on "computers in schools in Athens" the terms which identify the documents are: "computers", "schools" and "Athens".

The *a priori* relations are conceptual relations among terms that exist independently of any document and are generally recognized as such. In the above example, the term "computer" is related conceptually to "informatics", the term "schools" is related to "educational institutes" and "Athens" is related to "Greece". Any of these terms can be used to identify the corresponding documents. It is obvious that the *a priori* relations add a second dimension to the documentation and retrieval of information. The free text search is based on the *a posteriori* relations among terms in the sense that it uses terms that are automatically constructed from the full text of a document. On the other hand, the *a priori* relations are handled by a controlled vocabulary known as thesaurus, which incorporates both the terms and the relations among them.

Figure 1 depicts the distinction between the *a priori* and the *a posteriori* relations. The X axis shows the *a posteriori* while the Y shows the *a priori* relations.

From the aforementioned, it is obvious that the information retrieval process, which uses both search methods (free text and thesaurus), improves the effectiveness of searching, because it combines both dimensions.

The remainder of this paper is organized as follows: Section 2 presents basic concepts for a multilingual thesaurus, Section 3 describes our approach to extend the FDB (Frame DataBase) set-theoretic model, Section 4 presents the functionality of the proposed system, Section 5 illustrates the model with an example, Section 6 presents the advantages of our approach, Section 7 shows an evaluation of the proposed approach and Section 8 concludes the paper.

2. Adopting a Thesaurus System to Provide Conceptual Searches

Thesaurus is a vocabulary of standard terms of knowledge, selected from a natural language, organized in a specific hierarchy while the correlations between the terms are fully defined. The primary purposes of a thesaurus are indexing, saving and retrieval of data objects. Indexing is the process of assigning thesaurus terms to data objects. Retrieval is the process of locating data objects with the help of thesaurus terms [7]. A multilingual thesaurus is a controlled vocabulary selected from more than one natural language, specifies relations between terms as well as the equivalence terms in each of the languages chosen. Every relation is displayed and identified clearly by standardized relation indicators. A multilingual thesaurus allows definition of conceptual correlations/equivalences between terms selected from different natural languages.

Thesaurus terms can be classified into preferred and non-preferred. A preferred term (also known as descriptor) is selected through a set of equivalent terms to express a concept uniquely and consists of one or more words. A non-preferred term is an equivalent term which is not used for indexing and refers to the appropriate preferred term.

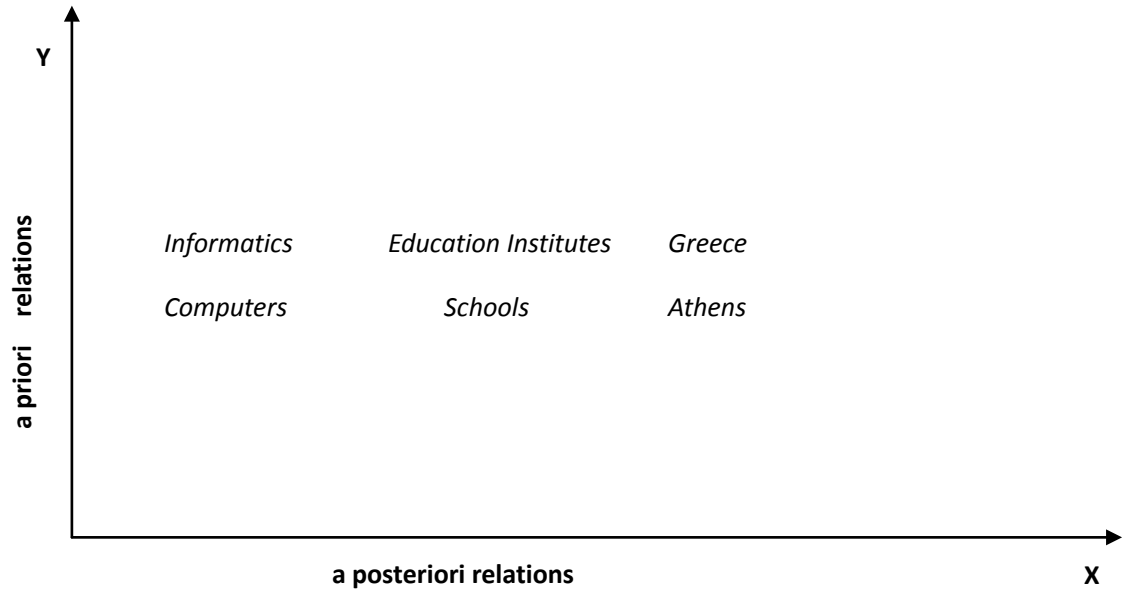

Figure 1. The *a priori* and the *a posteriori* relations

There are three kinds of fundamental relations that are regarded as language and culture independent: a) equivalence relationships that apply between preferred and the corresponding non preferred terms. They relate synonyms or quasi-synonyms to preferred terms (descriptors), b) hierarchical relationships that apply between preferred terms and relate descriptors to their broader or narrower terms, and c) associative relationships that apply between preferred terms to express concept proportions and meaning correlations (see Table 1).

Let us assume that T is the set of all thesaurus terms and t1, t2 two of these terms. These terms may be semantically related with one of the following relationships:

a) t1 is a term having wider meaning than t2 which means that t1 is a **Broader Term** (BT) of t2 (Figure 2a).

b) Conversely, t2 is a content with more specific meaning than t1 which means that t2 is a **Narrower Term** (NT) of t1 (Figure 2a).

c) Term t1 is not a fully broader term compared to t2 (Figure 2b).

d) t1 is a preferred term and it is **Used for** (UF) term t2 (Figure 2c).

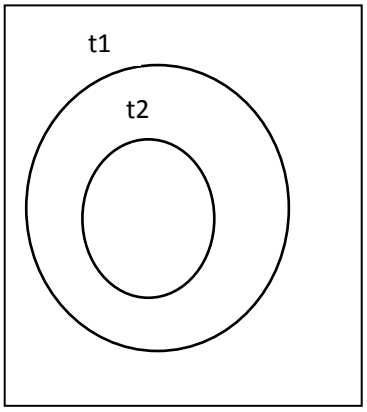

Figure 2a: Narrower and Broader terms

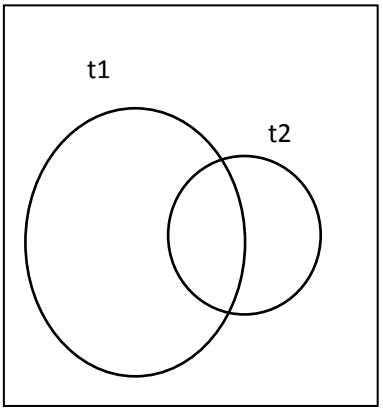

Figure 2b: Inexact equivalence

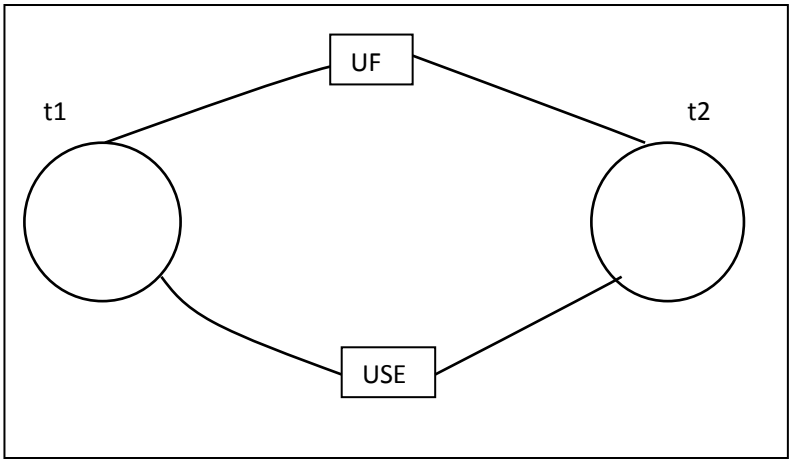

Figure 2c: Preferred and non-preferred terms

Table 1. Fundamental relationships between terms of multilingual thesaurus

Equivalence relationships	Hierarchical relationships	Associative relationships
USE UF (use for)	BT (broader term) NT (narrower term)	RT (Related term)

3. Extending the FDB Model

The FDB model allows the definition of any database by setting the appropriate metadata. It provides a universal schema and facilitates the definition of any database without any changes in the underlying schema. The current research extends the FDB model, so that its universal schema can be used to define one or more multilingual thesauri that can then form the basis for all subsequent searches. With the proposed changes the FDB model provides both traditional keyword search, as well as conceptual searches through the multilingual thesauri.

In this section we define all the necessary elements that extend the FDB model and introduce new objects in order to allow the handling of one or more multilingual thesauri. We also define appropriate metadata and introduce data store objects for a multilingual thesaurus. Finally, we present a practical example.

3.1. Multilingual Thesaurus Metadata and Data Store within the FDB Model

FDB is an integrated set-theoretic model for database systems that forms a framework for defining a structure (unified schema) that eliminates completely the need for reorganization at the logical level [4], [14]. FDB provides a universal schema which allows the definition of any database by the administrator who simply specifies the appropriate metadata. Amongst other utilities, FDB allows administration of multilingual databases both at data and interface level, definition of variable length objects (records in the traditional sense), etc. Any changes that may be necessary to be done to the database do not affect the universal database schema but simply concern the identification of the appropriate metadata. The basis for the creation of the unified schema is the definition and manipulation of metadata that compose the whole structure [5]. Accessing the information from an FDB schema becomes very easy with the use of simple statements provided by the Conceptual Universal Database Language (CUDL) [15].

Basic elements of the model are based on the mathematical theory of unordered sets and consist of the following sets: a) entities: the unordered set of registered entities that participate in the logical schema in our model, b) tags: the set of attributes describing each entity, c) subtags: the set of simple atomic attributes which constitute existing complex tags, d)domains: the set of all data domains, e) languages, vocabulary, messages: sets of strings or coded values that present human languages and corresponding messages [5].

In order to establish any data object in FDB we initially define the corresponding entity. Then, for each data entity we define the appropriate metadata (attribute tags) which concern: data elements (tags) that make up the data instance of each entity, the properties of each tag: e.g. occurrence status which defines mandatory or optional tags, repetition status which defines multi-valued or single valued tags, the data type of each tag, authority status which determines whether a tag references a tag of another entity or not. In FDB, the metadata denote the structure and the characteristics of each data object while separate objects are defined for data proper.

In order to accommodate one or more multilingual thesauri in FDB, it is necessary to define the appropriate metadata which are based on the universal FDB structure and concern the appropriate entities, tags and subtags. The entities to be defined are:

Metadata definition for multilingual thesaurus

Two basic entities are defined: one entity for the thematic terms of the multilingual thesaurus and another entity for the fundamental relationships that can correlate different thematic terms in a multilingual thesaurus:

a).**Thesaurus terms entity** (e_{thes}). This entity has a unique code and its basic attributes are designated according to the following attributes (tags):

- The saurus term tag which is mandatory, since the thematic terms must always be specified. This is a non-repeatable tag with a variable string data type.
- A class tag to denote whether a term is preferred or non-preferred. This tag is mandatory and has an authority status of "1" which means that it refers to another entity, the class entity.
- A status tag which is mandatory and denotes whether a term is authorized. Its authority status is set to "1" because this tag refers to a status entity.
- An optional tag for scope notes.

The authority status of class and status tags are set to "1" which means that other two entities exist in order to define all the classes and statuses of a term in any language. As for any other entity in FDB, the administrator can set any other attribute (tag) considered necessary for any thesaurus term.

b).**Fundamental thesaurus relations entity** (e_{rel}): This entity is used to identify all kinds of relations (Hierarchical, Associative and Equivalence) between thesaurus terms. It has a unique code and it is defined according to the following attributes (tags):

- A description tag which is mandatory and holds the description of the fundamental relationship.
- A short description tag, which is a mandatory tag used as a short description of a relation.
- A symbol tag which is mandatory and non-repeatable used as a symbol for a relation.
- Reverse relation tag, which is an optional tag that holds the code of the reverse relation. For example, the reverse relation of a NARROWER TERM relation is the BROADER TERM relation. When a new relationship between two thesaurus terms is defined, then the reverse relationship can be derived automatically.

c) **Class entity**: This denotes whether a term of a multilingual thesaurus is preferred (descriptor) or non-preferred (non-descriptor). This entity has one mandatory tag, the description tag.

d) **Status entity**. This entity is used to declare whether a term is authorized or not; only authorized terms can be classified as preferred or non-preferred and can be used afterwards to establish relationships. The status entity has one mandatory tag, the description tag.

The metadata defined above allow the identification of basic thesaurus entities: a) entity for multilingual thesaurus terms, and b) entity for determining the relationships between terms. Below are presented the objects added to the FDB model which can be used for storing the actual data of the multilingual thesaurus.

Definition of data store objects for the multilingual thesaurus

This section introduces the appropriate objects required for storing the data of a multilingual thesaurus which is similar to the objects used in FDB for each data record within the database. The following three new data store objects are required:

a) **TERM_DATA** object (see Appendix): used to store the terms of a multilingual thesaurus. Each thematic term is stored according to the specifications set by the corresponding metadata defined below for e_{thes} entity.

b) **RELATION_DATA** object (see appendix): used to store all fundamental relations that can correlate two thesaurus terms such as hierarchical, associative and equivalence relationships. Each relation item is stored according to the specifications set by the corresponding metadata defined below for e_{rel} entity.

c) **THESAURUS TERM RELATIONS** object (see appendix): holds two thesaurus terms and the relation that binds them.

The structure of the new objects is similar to the structure used in FDB for storing the data proper. This means that for each attribute (tag) we can hold data in many different languages, with a different repetition status, single or multi-valued tags, as well as the splitting of data into chunks for storing variable-length objects (records), where necessary.

Finally, an additional tag is introduced in any data entity which is necessary to achieve the association of a data record with one or more thesaurus terms. The authority status of the new tag is set to "1" which means that the tag is related with another entity, the thesaurus terms entity (ethes). The repetition status of the new tag is also set to "1" because each data record may be related with one or more thesaurus terms.

3.2. Architecture Diagram

Generally speaking, information retrieval from any relational database system is carried out by applying queries to the database via a query language such as SQL (Figure 3). A database user specifies keywords which form the search criteria that are applied to the database through well-structured SQL queries. In a traditional database system, all data records are checked; if one record contains these keywords then the record is returned into the result set since it meets the search criteria.

In FDB (Figure 3) the retrieval of database records can be performed using the information provided from one or more multilingual thesauri. The information retrieval process in FDB can be defined according the following steps:

- Firstly, the database user defines the search criteria through keywords which form the query.
- The keywords are searched in the thesaurus terms in order to identify all relevant and equivalent terms for all supported languages. Then the search criteria are enriched with all terms derived from thesaurus scanning.
- Searching of all the above terms in database records proceeds in two phases. Initially, the database user's keywords are searched into database records; if a record contains these keywords then it is returned into the first result set, say RS_1, of the query.
- The second result set is obtained by using the thesaurus terms; for each data record there is a tag which connects data records with thesaurus terms and holds the code of all the related terms. This gives a second result set, say RS_2, which is extracted using all relevant thesaurus terms (equivalent or related terms).
- The union of the above two sets $T= RS_1 \cup RS_2$ is the result set of the search and contains more relevant data records than the records retrieved from a traditional database management system.

In the proposed model, database users can choose whether they wish to apply information retrieval using a multilingual thesaurus or not.

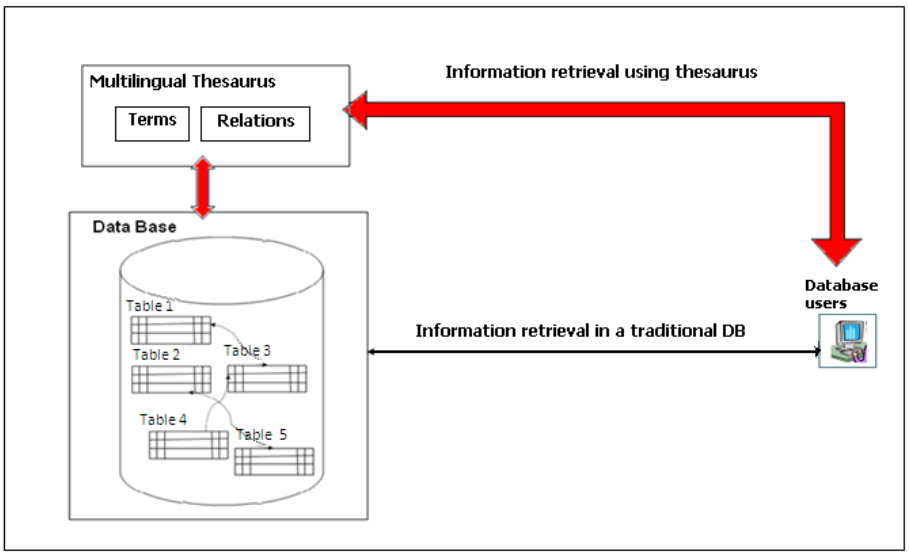

Figure 3. Information retrieval through a traditional DB and a multilingual thesaurus

4. Functionality of the Proposed System

The proposed system introduces a new dimension and represents an innovation in the field of databases at the level of searching and information retrieval from databases. The innovation lies in the fact that adding a multilingual thesaurus in the core structure of a database introduces a new layer on both information retrieval management and maintenance of the multilingual thesaurus. In what follows we present the functionality of the proposed system at multiple levels such as the innovations being introduced, the management of the multilingual thesaurus, the new dimension of the information retrieval process, the conditions and rules to be applied in order to ensure the integrity of the relationships between the data records and the multilingual thesaurus terms.

Structure of proposed system

- The proposed system allows the definition and the management of one or more different multilingual thesauri: for each new multilingual thesaurus that is being introduced in the database, a new entity is defined; thematic terms are inserted as well as the relationships between them. An extra tag is added to each data entity that provides the correlation between data records and thesaurus terms.
- More than one language can be supported for both database records and multilingual thesaurus terms: FDB supports multilingual data since the database administrator can define many different languages. Respectively, a thesaurus may be defined in FDB in all supported languages.
- The thesaurus administrator can define parametrically, in all supported languages, all different types of relations (such as equivalence, hierarchical and associative) which can correlate two thematic terms.

Management of thematic terms

- The thesaurus administrator can define the class, the status, the language and record scope notes and any other attributes required for each thesaurus term.
- For each type of relationship between thesaurus terms, the reverse relationship can be defined.
- For each type of relationship between thesaurus terms, a relationship symbol can be defined.
- Any thesaurus term can be correlated with others. Also, when a new relation between terms is established, it is possible to automatically create the reverse relationship. For example: when we establish that "ATHENS" is a NARROWER TERM of "GREECE" the reverse relationship can be automatically created by the system: "GREECE" is BROADER TERM of "ATHENS".
- An unlimited number of correlations are supported for each thesaurus term.
- It is possible to associate a thesaurus term in a language with the corresponding equivalent terms in one or more other languages.
- When establishing a relationship between thesaurus terms, the correctness of the relationship based on the status and class of terms can be automatically checked. For example, only preferred terms of a multilingual thesaurus can be involved in relationships, a non-preferred term cannot be correlated with other terms.
- When deleting a thesaurus term, we check whether the term participates in any correlations; in this case, the user is warned by the system, giving the option to either deny deletion, or delete the term with all its relationships.

Searching and presenting results

The architecture of the proposed system enables the utilization of domain independent thesauri which can support multiple monolingual or multilingual thesauri. As already mentioned, a multilingual thesaurus provides conceptual relations between terms; this knowledge offers a powerful search tool when retrieving records from a database. Below are features provided by the proposed system, when a multilingual thesaurus is used as a search-aid tool.

- When searching distinct keywords in the database, the result set will consist of records that include the specific keywords in all supported languages. For example, suppose that our search criteria include the keyword "Athens" and the multilingual thesaurus supports two languages: English and Greek. The result set will include records that contain the specific keyword (Athens) in English or in Greek. Any record with the keyword "Athens" or "Αθήνα" will be in the result set. Without the multilingual thesaurus, any record including the word "Αθήνα" would be excluded from the results.
- When non-preferred terms are in the search criteria (keywords) we can analyze the information provided from the multilingual thesaurus, retrieving records which include both non-preferred and corresponding preferred terms in all supported languages automatically. For example, suppose we look for the keyword "Αλεπουδέλλης" which is the non-preferred term of the authorized term "Οδυσσέας Ελύτης" (a well-known Greek poet). Using the information provided by the multilingual thesaurus, the preferred term "Οδυσσέας Ελύτης" will become a search term and all records including one or both of the two keywords will be in the result set.
- In every query applied to the database, it is feasible to perform this query using all equivalent terms as search criteria.
- In every query applied to the database the user can use the conceptual relationships derived from the usage forward and backward of the hierarchical structure of the multilingual thesaurus.

Reports production

A large number of reports can be obtained from the system, such as alphabetical lists of thesaurus terms, alphabetical lists of thesaurus terms with the relationships between them, etc.

5. A Practical Example

Let us adopt an example with three entities (see Table 2): a) entity with frame_entity_number 1, which is used to set multilingual thesaurus terms for a specific domain, b) entity with frame_entity_number 2, which is used to define all the

fundamental relations between multilingual thesaurus terms, and c) entity with frame_entity_number 100, which is used to hold data for books. Thesaurus terms and relations, as well as the book data, are multilingual. The supported languages are English and Greek.

Table 3 presents the basic attributes (tags) with the appropriate metadata for each entity of the entities in Table 2. For each attribute (tag) of an entity it is clearly defined whether it is optional or mandatory (occurrence status), whether it is a repeatable or no-repeatable tag (repetition status), whether it relates to a tag of another entity (authority status), and also its data type and length.

Table 2. Three example entities

ENTITIES		
Frame_Entity_number	**Title**	**Description**
1 (thesaurus terms)	1	
2 (thesaurus relations)	2	
100 (book)	10	

Table 3. Basic attributes (tags) of the entities in Table 2

TAG_ATTRIBUTES								
Entity	**Tag**	**Title**	**Occurrence**	**Repetetion**	**Authority**	**Datatype_id**	**length**	
1	200	20	M	N	N	1	50	(Term)
1	201	21	O	N	N	1	20	(DDC_Code)
1	202	22	M	N	Y	2	2	(class)
1	203	23	M	N	Y	2	2	(status)
……. Any other tag which may describe a thesaurus term								
2	300	30	M	N	N	1		(short description)
2	301	31	M	N	N	1		(symbol)
2	302	32	M	N	N	1		(description)
2	303	33	O	N	N	2		(reverse relation)
……. Any other tag which may describe a relation between thesaurus terms								
100	600	60	M	R	Y	1	50	(author)
100	601	61	M	R	N	1	512	(abstract)
100	602	62	M	N	N	1	512	(title)
100	**650**	**65**	**O**	**R**	**Y**	**2**	**1**	**(connection with one or more thesaurus terms)**
…….								

In the current example a multilingual thesaurus is defined for a computer science domain. According to the definition of tags, the object Term_Data (see Table 4) presents a few data of a multilingual thesaurus in two languages, English and Greek.

Table 4. Example Term_Data of a multilingual thesaurus

TERM_DATA						
Frame_entity_number	Term_Object_number	Tag	Lang	Repetition	Chunk	Data
1	1	200	EN	0	1	Computer Science
1	1	201	EN	0	1	004
1	1	202	EN	0	1	PREFERRED
1	1	203	EN	0	1	AUTHORIZED
1	2	200	EN	0	1	computer architecture
1	2	201	EN	0	1	004.22
1	2	202	EN	0	1	PREFERRED
1	2	203	EN	0	1	AUTHORIZED
1	1	200	GR	0	1	Πληροφορική
1	1	201	GR	0	1	004
1	1	202	GR	0	1	ΠΡΟΤΙΜΩΜΕΝΟΣ
1	1	203	GR	0	1	ΚΑΘΙΕΡΩΜΕΝΟΣ
TERM_DATA						
Frame_entity_number	Term_Object_number	Tag	Lang	Repetition	Chunk	Data
1	1	200	EN	0	1	Computer Science
1	1	201	EN	0	1	004
1	1	202	EN	0	1	PREFERRED
1	1	203	EN	0	1	AUTHORIZED
1	2	200	EN	0	1	computer architecture
1	2	201	EN	0	1	004.22
1	2	202	EN	0	1	PREFERRED
1	2	203	EN	0	1	AUTHORIZED
1	1	200	GR	0	1	Πληροφορική
1	1	201	GR	0	1	004
1	1	202	GR	0	1	ΠΡΟΤΙΜΩΜΕΝΟΣ
1	1	203	GR	0	1	ΚΑΘΙΕΡΩΜΕΝΟΣ

multilingual thesaurus in two languages, English and Greek.

Similarly, the object Relation_data sets out the fundamental relations that can correlate terms of a multilingual thesaurus (see Table 5).

Table 5. Example Relation_Data of a multilingual thesaurus

RELATION_DATA						
Frame_entity_number	Relation_Object_number	Tag	Lang	Repetition	Chunk	Data
2	1	300	EN	0	1	BT
2	1	301	EN	0	1	<
2	1	302	EN	0	1	Broader Term
2	1	303	EN	0	1	2
2	1	300	GR	0	1	ΕΥΡ
2	1	301	GR	0	1	<
2	1	302	GR	0	1	Ευρύτερος όρος
2	1	303	GR			
2	2	300	EN	0	1	NT
2	2	301	EN	0	1	>
2	2	302	EN	0	1	Narrower Term
2	2	303	EN	0	1	1

Correlations between thesaurus terms are presented in Table 6.

Table 6. Example correlations between thesaurus terms

THESAURUS_TERM_RELATIONS							
Frame_Entity_ Number1	Term_Object_ number1	Lang	Frame_Entity_ Number	Relation_Object _number	Frame_Entity_ Number2	Term_Object_ number2	
1	1	EN	2	2	1	2	(Computer science NT Computer Architecture)
1	2	EN	2	1	1	1	(Computer Architecture BT Computer science)

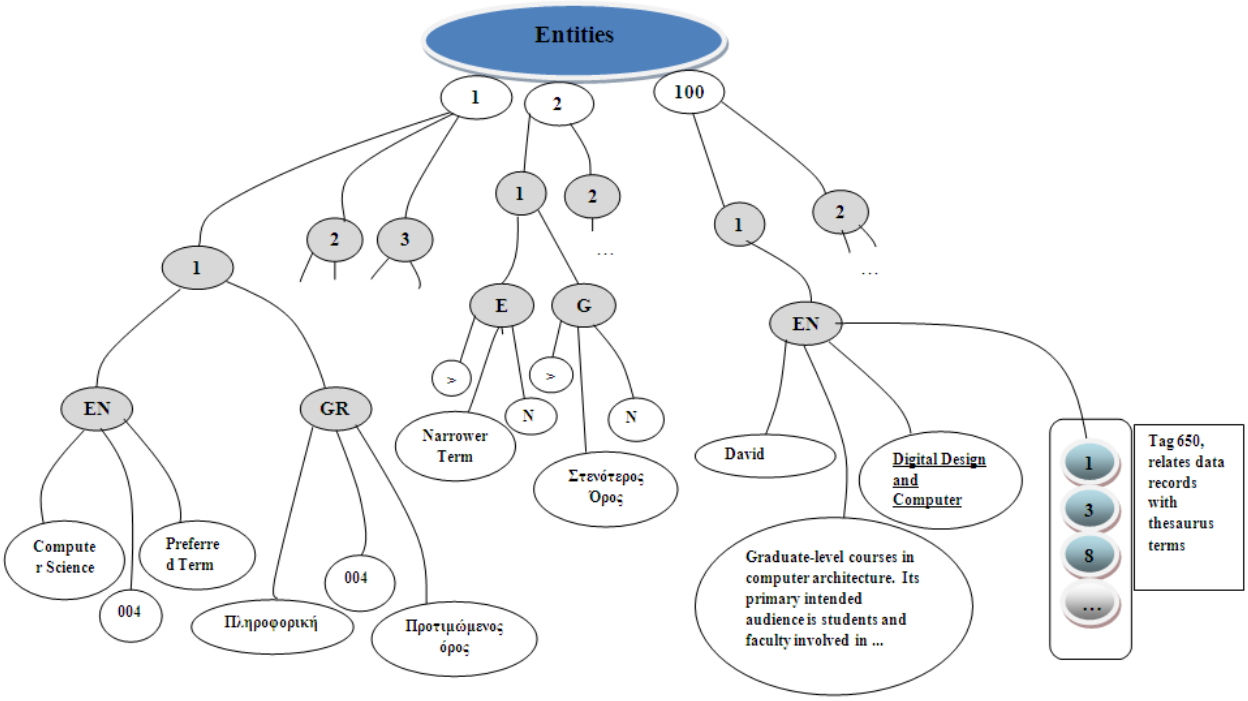

Figure 4. A practical example

Figure 4 shows a different representation of the example presented earlier, illustrating a data instance of the three example entities in two languages, English and Greek.

Table 7a. Example Tag_Data

TAG_DATA						
Entity	Frame_Object_number	Tag	Lang	Repetition	Chunk	Data
100	1	600	EN	1	1	David Harris
100	1	601	EN	1	1	Graduate-level courses in computer architecture. Its primary ...
100	1	602	EN	1	1	Digital Design and Computer Architecture
100	1	650	EN	1	1	1
100	1	650	EN	2	1	2

Table 7b. Example Authority_Link data

AUTHORITY_LINKS			
From_Entity	From_Tag	To-Entity	To_Tag
100	650	1	200

As already mentioned, each data row can be linked to many thesaurus terms. In the following table, the last two rows are used as links to associate the current book with the title “Digital Design and Computer Architecture” with two thesaurus terms: “computer science” and “computer architecture” (Table 7a). Authority links are used to define mappings between tags of different entities. For this example, the 650 tag of book entity is assigned to the 200 tag of multilingual thesaurus terms entity (Table 7b).

6. Advantages of Our Approach

FDB provides a universal schema, which allows the definition of any database, by setting the appropriate metadata. FDB also allows administration of multilingual databases, at both data and interface level. With the above extensions to the model, the accommodation of one or more multilingual thesauri is achieved. FDB is extended and becomes an integrated multilingual database management system that incorporates multilingual thesauri, and therefore allows the retrieval of data records using both free text techniques and conceptual searching using multilingual thesauri. The advantages of this approach are multiple:

a) Easy and flexible definition of any thesaurus: For each new thesaurus introduced, the administrator can define the terms that make up the thesaurus, as they can identify as many attributes (tags) required for each term and each relation without any restriction.

b) Compliance with ISO 2709 standard: a multilingual thesaurus can be defined as multilingual, according to the instructions set by the ISO 2709 standard.

c) Many multilingual thesauri can be established: the administrator can define one or more multilingual or monolingual thesauri. There are different thesauri for different domains. With the proposed structure, the administrator can define many thesauri, and database users will be in a position to utilize more than one thesaurus in their search queries. The establishment of multiple thesauri requires:

- The definition of a new entity for thesaurus terms with the respective attributes – tags (thesaurus metadata).
- Introduction of thesaurus terms (data).
- Definition of the correlations between thesaurus terms.
- Association of data records with thesaurus terms. This implies that an additional tag is required for each data entity, relating each data record with one or more thesaurus terms.

d) A new relationship, the "opposite term" relationship is defined: the introduction of the "opposite term" relationship allows the administrator to define for each thesaurus term the opposite concepts. All other relations, like *Broader Term*, *Narrow Term*, etc. are used during the data retrieval process to enrich searching with more keywords; this aims to identify more database records related to the search criteria set by the user. This new relation provides additional information about each term which, if exploited properly, can be used to reject irrelevant database records.

e) The proposed system enables database users, who issue queries, to choose whether they wish to use the information supplied by one or more multilingual thesaurus as search-aid data. This provides a powerful tool when retrieving data from multiple databases.

f) In many cases, for example bibliographic records, a database record may contain data in more than one language. These database records cannot be easily retrieved using a query that contains keywords from different languages. We suggest that queries containing keywords in more than one language can be structured using the information provided by a multilingual thesaurus, which is part of the database itself.

g) The proposed model presents an integrated environment that provides a universal schema which allows the definition of any multilingual database and also offers the ability to utilize one or more multilingual thesauri. The adoption of a search algorithm that takes into account all aforementioned parameters provides a powerful tool for retrieving multilingual information from a database. The proposed system constitutes an integrated database management system in which data record retrieval is carried out using both *a posteriori* relationships provided by free text and *a priori* relationships which are conceptual relations that exist independently and are generally recognized as such.

7. Evaluation of the Proposed Approach

The extended FDB model is an integrated system which allows the definition of any multilingual database and also offers the ability to utilize one or more multilingual thesauri. This section presents a study that evaluates the effectiveness of the multilingual search-aid thesaurus in terms of its effect on recall and precision.

7.1. The Test Database

In order to assure the credibility and validity of results, we chose a real environment for the selection and execution of test searches. The bibliographic database of the Athens University of Economics and Business was used for the following reasons:

- It contains 40,000 records of several material types (books, serials, papers, articles etc) documented using the international standard AACR2 and classified using DDC (Dewey Decimal Classification)
- It contains a multilingual thesaurus structured according to the international standard ISO 2709 [3] It consists of 20,000 terms approximately taken from Macrothesaurus and Eurovoc,
- The information retrieval software provides tools through the OPAC module (on-line public access catalogue) for searching using free text, thesaurus and a combination of both.

A number of frequently expressed queries were selected for testing proposes, which retrieved a wide range of records, using Boolean logic. Based on the multilingual thesaurus, the queries were formulated in both languages (English and Greek) and for each thesaurus relation. Starting from the basic search we evaluate the precision and recall of the records retrieved. Then, each query was expanded according to related and narrower terms forming the union search, the recall and precision of which were also evaluated.

7.2. Search-aid Multilingual Thesaurus

Because the main subjects of the bibliographic database are in the area of economics and informatics, we choose

appropriate thesaurus terms in order for the retrieved set of records to be large enough. According to the ISO 2709 standard [3], the terms we used are in the preferred form because the synonyms are considered non-preferred and are not used in the documentation. In the general case, synonyms can be used to improve recall. It is important to note that the multilingual thesaurus contains terms in two main languages, English and Greek and these languages are used in this in order to demonstrate the use of multilingual search.

Expanding free text searches using terms from multilingual thesaurus is considered a more effective method of searching because it involves both *a priori* relations which have been implemented in the thesaurus and the *a posteriori* relations identified by the free text.

7.3. Test Searches

The evaluation of the extension we propose can be conducted at different levels: a) at the logical level, which has to do with the relevance of the retrieved records, and b) at the level of performance, which refers to the retrieval time as far as the physical storage mechanisms are concerned. The latter case is not addressed by this paper.

We carry out logical level evaluation using the library system of our university which stores the queries formed by OPAC (on-line public access catalogue) users automatically. Twenty-five queries were selected for execution by the automated library system. The issue we investigated was the impact of the use of the search with a multilingual thesaurus as far as recall and precision are concerned. The search was made in several modes.

The *basic search* was the search formed by users without the search aid thesaurus. It is basic in the sense that it uses only the main concepts that were expressed by the users. An example of such search is: ECONOMICS AND MARKETING.

Using the search aid thesaurus, a number of queries can be constructed. In the narrower term search, every term/concept is matched in the thesaurus hierarchy and all the narrower terms are fetched. Without changing the overall structure of the basic search query, we extend it by disjunctions of those terms that were fetched. Continuing our example, the narrower term search would be:

(economics OR econometrics) AND (marketing OR prices)

Similarly, in the related term search the basic search was extended by including all the related terms found in the thesaurus for each basic concept, so the search would be:

(economics OR economist) AND (marketing OR advertising)

Combining all the above searches it's easy to construct a new query which includes all the terms, i.e.:

(economics OR econometrics OR economists) AND (marketing OR prices OR advertising)

Each term in the search aid thesaurus is expressed in several languages. This means that for each concept there are a number of hierarchies which express narrower and related concepts. Choosing the Greek hierarchy, we formed new queries in the same manner as above, starting with the *basic Greek search* and ending with the *union Greek search*.

Finally, we formed multilingual queries by combining terms from both languages. The multilingual basic search for the example above is:

(economics OR οικονομία) AND (marketing OR εμπορία)

while the multilingual union search is:

(economics OR οικονομία OR econometrics OR οικονομετρία OR economists OR οικονομολόγος) AND (marketing OR εμπορία OR prices OR τιμές OR advertising OR διαφήμιση)

Obviously, the above query will fetch all records fetched by the basic search as expressed by the library user and possibly a number of new records not included in the basic search. The degree of relevance of these records to the main concepts expressed above (namely *economics* and *management*) determines the usefulness of the search aid multilingual thesaurus.

For each query, the related and the narrower term relationships were not always both available. Moreover, the hierarchies in several languages may contain different number of terms for the same relationship. In both cases, we simply include only the terms that are available for each relationship.

Let B, BG, BM the sets retrieved by the basic search in English and Greek and the multilingual search respectively, and U, UG, UM the sets of union searches. Then: $U \supseteq B$, $UG \supseteq BG$, $UM \supseteq BM$ and $BM \supseteq (B \cup BG)$, $UM \supseteq (U \cup UG)$. It's worth mentioning that BM does not necessarily equal $(B \cup BG)$, because of the possible appearance of multilingual text in a record (the same applies also to UM and $(U \cup UG)$. Suppose for example that a record contains both the English word "economics" and the Greek word "εμπορία". Neither of the two searches will retrieve this record. It will only be retrieved by the multilingual search.

7.4. Search Result Evaluation

The records fetched from the database after executing the queries described above were tested for their relevance. A fetched record is considered relevant if expresses the concepts that were used in the basic search by the person who formed the query. Because this is subjective, the only person that can judge the relevance of the results is the person who submitted the query.

When a search is made in a language other than the language used to form the basic query, the terms used to form it may have several meanings in the specific language. This may lead to a number of retrieved documents that are completely irrelevant. Thus, it's interesting to investigate the impact of the search aid multilingual thesaurus in the precision of the results.

As far as recall is concerned, it's well known that it is difficult to estimate it in large databases. On the other hand, what we are really interested in is to compare the different values of recall for each query formed by using the search aid

thesaurus. So, instead of estimating the absolute recall of the results, we compute the relative recall of each query assuming that the recall of the multilingual union search is 100% [13]. In other words, we assume that the only relevant documents for a query that exist in database are those retrieved by the multilingual union search. The recall of each of the other search modes is then computed as the percentage of the relevant documents in the total number of documents retrieved by the union search. This is because every record retrieved by each search mode is also retrieved by the multilingual union search mode.

7.5. Findings – Estimations of Relative Recall and Precision

The search outputs contain the total 4475 records of which 2339 were relevant and 2136 were non-relevant. The number of retrieved records for the above forms ranged from 3 to 2875. This significant difference in the retrieved records is due to disjunctive form of some basic queries and to the use of the use of the multilingual thesaurus, which enriches the basic form with alternative concepts in both languages.

Comparing the recall and precision of the basic and the multilingual union search of the 25 queries shows the total effect of the multilingual search-aid thesaurus. The recall and precision were estimated using the formulas:

$$R = \sum_{i=1}^{n} \frac{r_i}{tr_i} * \frac{100}{n}$$

$$P = \sum_{i=1}^{n} \frac{r_i}{td_i} * \frac{100}{n}$$

where R and P are the recall and precision respectively, n is the number of queries, r_i is the number of relevant and retrieved records, tr_i is the total number of relevant records, td_i is the number of retrieved records (relevant and non-relevant) for the i_{th} query. We assume that the multilingual union search retrieves all the relevant records that exist in the database resulting in the relative recall of this search being 100%. Based on this hypothesis, the relative recall of the basic search is 44.5%, only about half of the relative recall of the multilingual union search, demonstrating the total effect of the multilingual search-aid thesaurus.

The precision of the basic search is 59.2%. Compared to the 50.4% which is the precision of the multilingual union search, there is only 9% decrease as the result of using the thesaurus (see table 8).

The multilingual search-aid thesaurus affects the relative recall and precision in two distinct ways:

a) Enrichment of the basic search with alternative concepts (using narrower and related-synonym terms),

b) Enrichment of the basic search using equivalent terms in other languages (in this study Greek terms). This is shown in Table 9, which presents the relative recall and precision for each of the five forms of search.

The inclusion of narrower and related synonym terms increases the relative recall from 44.5% in the basic search to 65.4 in the union search, while the precision decreased only from 59.2% to 52.1%. For the Greek search, the relative recall increased by 10%, while the precision decreased by 21.2%. This considerable decrease in precision is due to the fact that terms in different languages may express other concepts, in addition to the basic concepts. For example, an equivalent term for the English word “management” is “διεύθυνση” in Greek, which expresses many distinct concepts such as “address” and “administration”

The multilingual capabilities of the search-aid thesaurus increased the relative recall of the union search by 35%. This increase can be attributed to the Greek terms which were added to the union search, forming the multilingual search. It’s important to note that the set of records retrieved by the multilingual search exceeded the union of the retrieved records of the two union searches (English and Greek) by 120 records, which is due to multilingual data, as explained in section 7.3.

Table 8. Relative recall and precision for the basic and multilingual search forms

Search form	Relative recall Average %	Relative recall SD	Precision Average %	Precision SD
Basic search	44.5	29.1	59.2	16.1
Multilingual union search	100	0	50.4	15.7

Table 9. Relative recall and precision for five search forms

Search form	Relative recall Average %	Relative recall SD	Precision Average %	Precision SD
Basic search	44.5	29.1	59.2	16.1
Basic Greek search	20	16.4	61	26.1
Union search	65.4	24.5	52.1	18.8
Union Greek search	30	18.7	39.8	11.6
Multilingual union search	100	0	50.4	15.7

Figure 5 shows the histogram of the values of relative recall and precision for the five forms of search.

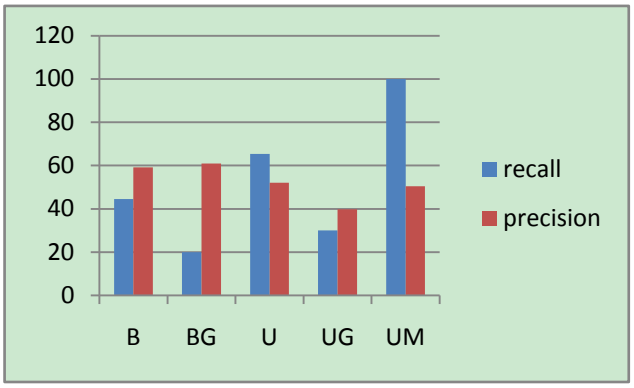

Figure 5. The values of relative recall and precision

7.6. Statistical Significance of the Findings

The difference in recall and precision between the basic and the multilingual union search, as well as between the union and the multilingual union search were tested for the statistical significance. For this purpose, a non-parametric test was used because the sample size (25 queries) was small and the distribution of the population was not known. Of the non-parametric tests, Wilcoxon matched-pairs signed-rank test was chosen because the study employs related samples (the records fetched by the multilingual union search incorporates all the records fetched by the other search modes) [6].

It was found that the increase in relative recall in both cases was statistically significant with a significance level of 0.01. This means that the improvements achieved using the multilingual search-aid thesaurus is considerable. The decrease in precision on the other hand is statistically significant in the case of the basic (both Greek and English) and the multilingual union search (a=0.01), but was not found to be statically significant in the case of the union and the multilingual union search. This means that the use of multilingual relations does not decrease precision considerably.

8. Conclusion and Further Research

This paper presented an innovative approach concerning conceptual retrieval through multilingual thesauri, which are integrated with the FDB model. The FDB model provides a universal schema which can hold any database schema whereby changes that occur at the logical level do not affect the universal schema; instead, the FDB structure accommodates changes within the corresponding metadata. In a traditional database system, information retrieval takes place through free text search techniques. With the proposed extensions to the FDB model, the usage of a multilingual thesaurus mainly concerns the information retrieval process from database and enables a database user to apply conceptual queries to the database. The information retrieval process utilizes both *a posteriori* relationships, denoted through free text search, and *a priori* relationships, which are conceptual relations between terms that exist independently of any database. The *a priori* relationships are provided by a multilingual thesaurus.

In the proposed model the multilingual thesauri form part of the general database system which means that they are exploited by default in order to carry out conceptual retrieval and correlation of records. Our approach offers a very powerful tool for all applications, which can therefore concentrate on the services offered to users rather than on defining and maintaining complex search strategies. Note that traditional approaches usually use the thesaurus as an independent utility which is simply linked to the existing database software and is thus application-dependent.

Future work involves the definition and formal representation of the search algorithms. Search algorithms in FDB should enable concrete search features for the database user and exploit the information provided from multilingual thesauri in a combinatorial manner automatically. Another research direction is the generation of the thesaurus trees automatically and the population of the corresponding entities of the database. This issue lies in the field of Natural Language Processing (NLP). Finally, future work should also address the usage of the multilingual thesaurus as a search-aid tool in a real-time environment in order to draw statistical assessments at the conceptual level.

REFERENCES

[1] C. Kapetis, Multilingual Thesaurus Automated System: A dynamic tool for information retrieval and documentation, 1st Greek Technical Chamber Conference: Info Society, Athens 4-6 December 1995, pp. 591-597.

[2] ISO 2788, Documentation-Guidelines for the establishment and development of monolingual thesauri, 1998.

[3] ISO 2709, Documentation-Guidelines for the establishment and development of multilingual thesauri, 1987.

[4] E. J. Yannakoudakis, P. K. Andrikopoulos, A set-theoretic data model for evolving database environments, In Proceedings of the International Conference on Information & Knowledge Engineering, IKE 2007, June 25-28,2007 Las Vegas, Nevada, USA.

[5] E. J. Yannakoudakis, An efficient file structure for specialised dictionaries and other 'lumpy' data, International Journal of Information Processing & Management, Vol. 23, No. 6, pp. 563-571, 1987.

[6] Siegel, S., Castellan, N.J. , Non parametric statistics for the behavioral sciences, New York: McGraw-Hill, 1988.

[7] R. Kramer, R. Nikolai, C. Habeck , Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies, International Journal on Digital Libraries, Springer – Verlag 1997.

[8] S. Gauch, J. Wang, and S. M. Rachakonda, A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases, ACM Transactions on Information Systems, Vol. 17, No 3, Pages 250-269, July 1999.

[9] C. R. Watters, Logic Framework for Information Retrieval, Journal of the American Society for Information Science, 40(5): 311-324, 1989.

[10] Fang Liu, Clement Yu, Weiyi Meng, Abdur Chowdhury, Effective keyword search in relational databases, In Proceedings of the ACM SIGMOD international conference on Management of data, pp. 563-574, 2006.

[11] Bhalotia Gaurav, Hulgeri Arvind, Nakhe Charuta, Chakrabarti Soumen, Keyword Search in Databases, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 3, pp. 22-32, September 2001.

[12] Vagelis Hristidis, Yannis Papakonstantinou, Discover: keyword search in relational databases, Proceedings of the 28th international conference on Very Large Data Bases, pp. 670 – 681, 2002.

[13] Kristensen J, Expanding end-users' query statements for free text searching with a search-aid thesaurus, Information processing and management, Vol. 29, No. 6, pp. 733-744, 1993.

[14] Yannakoudakis E.J., Tsionos C.X. and Kapetis C.A, A new framework for dynamically evolving database environments, Journal of Documentation, Vol. 55, No. 2, pp. 144-158, 1999.

[15] Yannakoudakis E. J., and Nitsiou M., A new conceptual universal database language (CUDL), In 2nd International Conference From Scientific Computing to Computational Engineering, Athens, Greece , 2006.