SIS - TMS : A Thesaurus Management System for Distributed Digital Collections

Martin Doerr[1] and Irini Fundulaki[1]

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas, Science and Technology Park of Crete, Vassilika Vouton, P.O. Box 1385, GR 711 10, Heraklion, Crete, Greece
{martin, fundul}@csi.forth.gr

Abstract. The availability of central reference information as thesauri is critical for correct intellectual access to distributed databases, in particular to digital collections in international networks. There is a continuous raise in interest in thesauri, and several thesaurus management systems have appeared on the market. The issue, how to integrate effectively such central resources into a multitude of client systems and to maintain the consistency of reference in an information network has not yet been satisfactorily solved. We present here a method and an actual thesaurus management system, which is specifically designed for this use, and implements the necessary data structures and management functions. The system handles multiple multilingual thesauri and can be adapted to all semantic thesaurus structures currently in use. Consistency-critical information is kept as history of changes in the form of backward differences. The system has been installed at several sites in Europe.

1 Introduction

Modern information systems typically consist of a number of autonomous information sources and provide access to huge amounts of heterogeneous information. To overcome the difficulties of handling a multitude of heterogeneous interfaces and data structures, free-text search engines are widely used. They provide access under least assumptions and hence are ultimately limited in precision and recall. In particular they do not solve per se the problem of appropriate terminology and multilinguality for search request formulation.

Besides database federations with integrated schemata, the use of thesauri and other kinds of reference information, so-called "authorities", have been successful means to improve access to "verbose" (texts, [1], [2], [3], [4]) and "non-verbose" data (images, data records etc. [5] [6]) residing in multiple heterogeneous and possibly multilingual information sources. (By „verbose“ we mean a text large enough such that one can sufficiently conclude with statistical methods on its contents).

C. Nikolaou, C. Stephanidis (Eds.): ECDL'98, LNCS 1513, pp. 215-234, 1998.

One may distinguish three kinds of use of authorities (compare [7], [8], [9]):

- Guide the user from his/her naïve request to the use of an set of terms optimal for his purpose and for the characteristics of the target information source. If the respective knowledge about the target is not explicit or implicit in the thesaurus, the results are limited. If the targets are many, it is not practical.
- Expand naïve user terms or the terms optimal for the purpose of the user into sets of terms optimal for each different information source. In the case of full-text retrieval, this may mean to produce a weighted list of possible common words for the user's concept. For structured database queries, this may mean to select only the closest term in use on that system [10]. Still the degree of matching between the requested terms and the used terms are undefined.
- Classify all information assets of a certain collection with controlled vocabulary from a specific thesaurus. Together with the above measures., a well defined matching of query terms and target terms [10] can be achieved.

The above holds for monolingual sources using the same or different thesauri, or multilingual sources, as long as appropriate transitions, translations or correlations can be established. The focus of this paper is to present methods and an actual system suited to store, maintain and provide access to knowledge structures that are in use or needed for these three tasks and the respective auxiliary system interfaces.

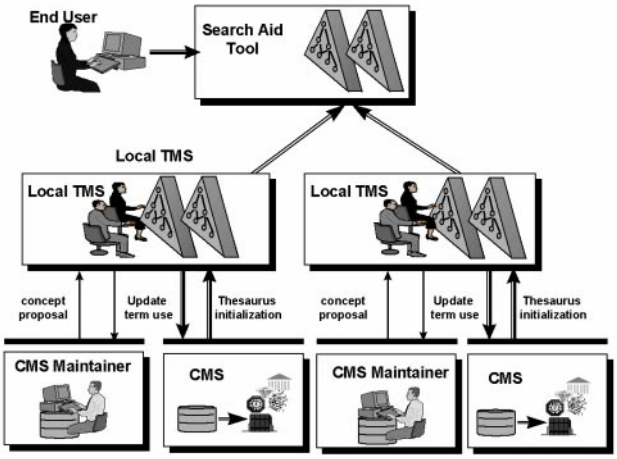
The creation of multilingual thesauri is a crucial problem in this context, because traditional human editing is extremely labor-intensive [11]. We restrict ourselves in this paper to the support of this process by a central data management system. The application of statistical methods, computational linguistic methods, elaborate CSCW means and their combination for efficient production of multilingual thesauri are subject of the recently started Term-IT project (TELEMATICS project LE4-8356).

The simultaneous and remote access to hundreds or even thousands of target systems requires the complete integration of thesaurus tools in a wide area information environment. For optimal results, the terms used for asset classification, in the search aid thesaurus and in the experts' terminology should be consistent. This led us to the vision of a three level architecture of components cooperating within an information environment: vocabularies in local databases, local thesaurus management systems of wider use and central term servers for retrieval support**.**

Typically, local databases have a more or less idiosyncratic way to enforce vocabulary control. For reasons of standardization of format and centralization of handling, we foresee an independent thesaurus manager to which the vocabularies of several local databases are loaded, and in the sequence organized as thesauri ("authorities") by an expert, following variations of the ISO2788 semantic structure. In addition, standard external vocabularies can be loaded. These authorities may be specific to one database, a user organization, or a whole language group. The local vocabularies and terms already used for classification may need updating with changes done at the thesaurus manager.

Search agents or user interfaces for information retrieval need knowledge of the authorities in local use, at least of the higher level terms, or knowledge about the language in case of free text search. Therefore they must communicate with one or

more term servers, which hold-released versions or extracts of the local authorities. Moreover, a term server must be fed with equivalence expressions between the meaning of terms in different authorities, either by an expert team or by linguistic methods and subsequent human control. These expressions are used to replace the terms in a user request with more or less equivalent terms of the target system - automatically or in a dialogue with the user. As equivalence expressions are difficult to produce, term servers containing different translations may be cascaded to make multistep replacements, e.g. Finnish to English, English to Greek, etc. Of course



precision will suffer.

Fig. 1. 3-level terminology service

This three stage architecture ideally reflects the practice and needs of classification, expert agreement, user organization and search aids. It is a fully scalable solution which has already been partially realized in the AQUARELLE [12], [13] project and system. The methods and system development we present in the following are targeted to be components of such an environment and architecture. The work presented here is outcome of a series of cooperations with libraries, museums and cultural bodies, but of other domains as well.

1.1 Thesaurus Management Systems

The use of thesauri for classification and information access in global information spaces becomes complex due to the large number of existing or needed thesauri; a fact that is justified by existence of multiple thesaurus providers from different scientific backgrounds with different practices, aspects and history. Multi-thesaurus systems have been proposed in order to provide access to information sources in global spaces classified with multiple and possibly interconnected thesauri.

The basic research issues associated to multi-thesaurus systems are
- the maintainance of the autonomicity and independent evolution of the component vocabularies
- the incorporated methodology for the treatment of interthesaurus links and
- the role of the system in a retrieval process.

[8] and [11] propose the construction of a multi-thesaurus system that incorporates multiple interlinked thesauri. [8] proposes SemWeb, „an open multilingual, multifunctional system for integrated access to knowledge about concepts and terminology". A 3-level architecture is envisioned by the author: the incorporated sources contain terminology from thesauri and other kinds of authority data; a common interface that provides a gate-way to the sources and an evolving and integrated knowledge base that incorporates terminology from the underlying sources, that are able to maintain their autonomy.

Under the same perspective, the authors in [14] propose a thesaurus federation that draws upon mediation [15] as the technique for database integration. In the mediator, metadata of the underlying vocabularies are stored in a repository. The actual thesaurus contents are not stored in the mediator but the integration of thesauri is perceived with the incorporation of the existing interthesaurus relations in a term mapper module, allowing the component thesauri to maintain their autonomy.

In the „Vocabulary Switching System" [16] existing thesauri are incorporated in the system but no interthesaurus relations are established. [17] presents an example of this approach with agricultural thesauri. Although the component thesauri are able to maintain their autonomy and evolve independently, there is loss of recall during the retrieval process. The retrieved items are only those indexed with the thesaurus used for retrieval, and in order to ensure recall, either the user must manually incorporate search terms from other thesauri, or the underlying sources must be classified in multiple vocabularies, a time consuming process that requires extensive knowledge of the incorporated terminologies.

Authors in [18] propose a method for the construction of an interlingua to reduce the established mappings between the component thesauri in a multi-thesaurus system to about one per term.

Merging thesauri is another approach to the construction of a multi-thesaurus system. During a merging procedure, a single thesaurus is built from a set of others. Terms referring to the same concept are identified and federated into a unique concept, and consequently all inconsistencies are identified and resolved either automatically or semi-automatically. Work in thesauri merging has been performed by [19], [20], [21] and [22]. Merging is only appropriate for a homogeneous environments because of the intellectual and coordination problems it raises.

In [23] the authors point out that the difficulty of locating the appropriate terms hinders query formulation and indexing of a collection. A number of thesaurus management systems offer only flat lists of selected terms from which the user can choose one, which are not ideal if the number of choices is large. Direct access to the contents of a thesaurus is also provided, but this is considered not to be useful to users

who are not aware of the structured vocabulary. In addition to the above, navigation in a hypertextual format is supported with the use of the broader/narrower relations of terms that can be proved slow if multiple levels of the hierarchies must be traversed to access the desired terms. A rather interesting work on a hypertextual interface is presented by the authors in [24]. The user can move at will in a hypertextual ordered space, selecting terms for immediate use or for subsequent searches without leaving the thesaurus or going into separate search mode.

A number of commercial thesaurus management systems have been produced to assist users in the development of structured vocabularies. Some have been developed as modules of complete indexing and retrieval systems while others can be acquired and used independently of any software. Nearly no one has a client-server architecture. Some of the most interesting thesaurus management systems are the Thesaurus Construction System (TCS)(http://www.liu-palmer.com/), the MultiTes (http://www.cris.com/~multites), the STRIDE(http://www.questans.co.uk/),the STAR/Thesaurus(http://www.cuadra.com/) and the LEXICO/2 systems (http://www.pmei.com/lexico/lexico.html). To our knowledge the majority of the existing commercial thesaurus management systems are compatible with the ANSI/NISO standards, support the evolution of multiple monolingual and multilingual vocabularies (TCS, MultiTes), provide consistency mechanisms to check for reciprocal relations, duplicate terms and non-consistent cross-references (MultiTes, LEXICO2) and support batch and/or interactive editing of thesaurus contents. An interesting feature of some of the systems such as STRIDE, and the STAR/Thesaurus is the maintenance of a log of the transactions, although it is not completly clear how this information is represented and eventually stored. Another interesting feature is the ability to create user defined relations (MultiTes). Although this feature adds great flexibility to the system, the absence of consistency control on the created relations creates difficulties in their maintenance. Most of the systems support many formats for the representation of information, mostly textual and/or hypertextual and only few provide graphical interfaces.

2 Requirements for Thesaurus Management

We roughly divide the requirements into those for (1) interaction with the thesaurus contents except manipulations, (2) maintenance, i.e. the manipulation of the contents and the necessary and desirable support of associated work processes, and finally (3) analysis, i.e. the logical structure needed to support (1), (2), and the thesaurus semantics in the narrower sense.

Interaction needs. The interaction needs with the thesaurus contents split into man-machine interfaces and interfaces to other systems. Most literature concentrates on the first point. In [9], [11] we have pointed out the relevance of system interfaces as well. Besides literature and experience with users of our systems, we refer here to [25] and related user meetings organized by the Museum Documentation Association, UK, the Getty Information Institute, CA, and the AQUARELLE project [26].

The most prominent human interaction is the identification of concepts for retrieval purposes. Without going into details, strategies can be:

- Linguistic - search by noun phrases, words, part of words or misspelled words. Often however similar concepts do not have terms of any linguistic similarity, e.g. "knife" and "dagger".
- Hierarchical – search by narrowing down from broader notions of the same nature, e.g. "sword" –> "foil", or navigational in nearby branches as "red" – "pink (color)", "small" – "large".
- Associative – search by related notions or characteristic context, e.g. "bridges" – "bridge construction", "baby" – "dolls".

The next step is the understanding of the concepts and matching against what he/she had in mind. As well, he/she has to verify, if the concepts found are the best choice within the given vocabulary. Understanding is supported by explanations ("scope notes"), multimedia examples, or schematic presentations and the term environment of associative and hierarchical links.

Effective interaction is a problem of presentation and of semantic analysis. Most views should be available as text and graphics. Semantic views must be very flexible. They must render good overviews of local environments which show many kinds of relations, and "global" overviews which show few kinds of semantic relations at a time. Disorientation must be avoided during navigational access, e.g. by global views and by logging previous steps/paths. Furthermore, interaction speed is crucial, as users may easily give up if response is too slow.

Last, construction, maintenance and quality control of thesauri requires a different kind of access, characterized by global views on certain properties, missing properties, conflicting declarations, statistics and transaction information. Questions of completeness and clear distinction between concepts are important as well.

System interfaces may serve the following purposes:

- Enabling term browsing as search/classification aid from within an external GUI.
- Automatic term expansion/translation for retrieval mechanisms.
- Term translation/replacement in order to update obsolete classification terms in some information base.
- Term verification for vocabulary control within an external application.

This requires an API and a client-server architecture. As thesauri are central resources, the capability to communicate on low bandwidth over WAN is important. The API functions needed are very simple. Up to now, such interfaces are subject to customization, but they could easily be standardized and make the use of thesauri in actual application significantly easier. Interfaces of the first two types have been developed and experimented with in the AQUARELLE project. The authors currently engage in experimentation with the other interface types.

Maintenance needs. One can roughly distinguish between maintenance of semantic structure and workflow support. Semantic structures have primarily to be maintained by interactive data entry facilities, that easily allow to manipulate hierarchical structures and other links, preserving constraints as connectivity, referential integrity, anticyclicity etc.

Market systems and most literature so far do not deal with the problems that arise, when a thesaurus is developed outside of the databases which use its terms for classification. The ISO2709 has a link for the case, when a concept is split into siblings. Such isolated semantics do not provide a solution. The basic problem is that vocabularies of hundreds of thousands of terms cannot be compared by hand with millions of data records in order to migrate to the new edition.
There must be a notion of a release and the effective changes between releases as analyzed in the following chapters. The current practice to backtrack modification dates for that purpose is not satisfactory.
The process of gathering concepts from users and experts, quality control and their embedding in a large thesaurus can be rather complex. Distant users and experts need to communicate and make thousands of small agreements [11]. Much work and research is needed on this field, which is out of the scope of this paper.

Requirements of Analysis. As follows from the above, the conceptual model of a thesaurus management system needs a great flexibility. All modern systems allow for adding user defined semantic relations. To our opinion, this is not enough. Rather all of the following aspects have to be considered and made configurable to a certain degree:

- logical and linguistic links within and between thesauri, as BT, UF, equivalence etc. (see below), which are relevant for query processing
- rules for dynamic concept formation (e.g. combine "factories" with process terms)
- context associations to assist browsing (e.g. the ISO2709 "subdivisions", in which database field the term is used, how many items it classifies).
- explanations as scope notes, , source references and multimedia [27] for human understanding
- migration information between releases
- workflow information about proposals, decisions, term status, persons and groups involved, following steps, "todo" etc.

Obviously, this is open ended, and we shall describe in the following which choices we have made in our systems for the time being.

3 The SIS-TMS

The SIS-TMS is a multilingual thesaurus management system and a terminology server for classification and distributed access to electronic collections following the above analysis. The its distinct features are its capability to store, develop, display and access multiple thesauri and their interrelations under one database schema, to create arbitrary graphical views thereon and to specialize dynamically any kind of relation into new ones. It further implements the necessary version control for a cooperative development and data exchange with other applications in the environment.

It originates in the terminology management system (VCS Prototype) developed by ICS-FORTH in cooperation with the Getty Information Institute in the framework of a feasibility study. It was enhanced within the AQUARELLE project, in particular by

the support of multilinguality. An earlier version is part of the AQUARELLE product [26]. A full product version will be available summer '98.

The SIS-TMS is an application of the *Semantic Index System* (described below), a general purpose object-oriented semantic network database, product of the ICS-FORTH with client-server architecture.

3.1 The Semantic Index System

The Semantic Index System (SIS) [28], a product of the Institute of Computer Science-FORTH, is an object oriented semantic network database used for the storage and maintenance of formal reference information as well as for other knowledge representation applications. It implements an interpretation of the data model of the knowledge representation language TELOS [29] omitting the evaluation of logical rules. A formal treatment of this data model can be found in [30].

The structures and the modeling constructs of SIS allow the representation of complex thesaurus structures in an elegant and compact way. SIS attributes referring to entities are implemented as bi-directional, directed, typed links. We model terms, descriptors, persons, sources etc as entities. Hence for all of them vocabulary control and referential integrity is enforced automatically throughout the system. There is no need to make special arrangement for multi-valued relations, as broader terms, related terms, synonyms etc. Multiple instantiation is used besides others to organize terms by semantic and administrational criteria, to organize workspaces and to view multiple thesauri once together, once separate. Metamodels are used to annotate consistency constraints to be enforced by the application. The dynamic schema of SIS allows a graceful evolution of thesauri into richer knowledge bases. Finally, the system is highly optimized for fast referential access.

3.2 Thesaurus Structures

Assumptions on Concepts. According to [7] one of the major purposes of a thesaurus is to "provide a map of a given field of knowledge, indicating how *concepts* or ideas about concepts are related to one another, which helps an indexer or a searcher to understand the structure of the field.

We distinguish concepts from terms, in contrast to IS2788. Cognitive scientists have proposed several definitions for the notion of "concept" (e.g. [31]). According to one point of view, a concept is perceived as a set of entities, called "concept instances" characterized as such by common agreement rather than formal reasoning on the properties that characterize an individual entity as an instance of a concept. We adopt this view for thesauri, considering a concept as a notion by which some people agree to refer in a well defined manner to a set of real world objects with the same properties, without necessarily defining properties. Consequently, certain semantic relations between concepts are interpreted as relations between sets as will be presented below. For more details see [22], [10].

Following ISO2788, we regard terms as nouns or noun-phrases, by which groups of people use to refer to certain concepts in a certain context. Due to varying groups and contexts, concepts and terms are related many to many.

Modelling Thesaurus Notions. The SIS-TMS schema is extensible at run-time. New semantic relations can be created or existing ones can be specialized. The current conceptual model of the SIS-TMS for the representation of multiple interlinked thesauri incorporates the thesaurus notions and intrathesaurus relations of the ISO2788 for monolingual thesauri and an extended version of the ISO5964 interthesaurus relations [10]. In prototype versions, this schema has been extended for the ULAN and TGN, vocabularies of the Getty Information Institute and the Library of Congress Subject Headings. We mainly use ISO2788 terminology for the names of the classes and relations in the SIS-TMS schema. In the manner of semantic networks, these names are directly presented in the user interface together with the respective data and read quite naturally (See fig. 5, left window).
We model *Preferred Terms* for indexing and *Non-Preferred Terms* as synonyms and entry points for the user. In addition, *Non-Preferred Terms* may be used for full-text retrieval. We adopt the notion of *Descriptor* of the Art & Architecture Thesaurus [34] according to which: "a descriptor is the term that uniquely identifies the concept". Hence a *Descriptor* is a term and a concept identifier in double nature. All other terms, preferred or not, are related to the concept and not further described, as we are not interested in linguistics.
As the *concept* is identified by a *descriptor*, i.e. by a linguistic expression that best expresses the common understanding of experts or public and it must be unique within the context in which it has been defined, it may not be exactly the word an expert uses. For instance, "pink (color)" and "pink (vessel)" would be good descriptors, but experts would say "pink" in both cases. In the SIS-TMS, all terms and descriptor names are enforced to be unique throughout the database. Terms may be multiply related to different concepts (*Descriptors*), but if a good term appears to conflict with a descriptor, the descriptor has to be renamed, i.e. usually extended for disambiguation.

Concepts carry all the intra and interthesaurus relations that make up the semantic structure of the thesaurus contents, they carry the administrational information, and they can be described by scope notes and understood language independently.

Figure 2 shows the isA hierarchy of the *SIS-TMS* Classes of thesaurus notions. We use in the following "abstract" for classes which are not directly instantiated, and "abstract hook" for abstract classes, which are designed to be superclasses of classes in future extensions. „ThesaurusNotion“ is the abstract root. „ThesaurusExpression“ is the abstract hook for terms, person names, date expressions etc. „ThesaurusConcept“ is the abstract hook for concepts in the above sense, persons, places etc. „HierarchyTerm“ is the class for concept in the above sense, those that can be generalized or specialized into broader/narrower meaning. It combines *Node Labels*, or "guide terms" and descriptors. We do not distinguish functionally between both (see e.g. [32]). „AlternativeTerm“ is the complement of „Descriptor“. „Topterm“ are those having no broader terms. „ObsoleteDescriptor“ are abandoned concepts

(sometimes thesauri decrease, e.g. in favor of dynamic concept formation) and finally „ObsoleteTerm“ are deleted noun-phrases. The latter two serve version management for referential integrity incremental update.

Intrathesaurus Relations. The semantic relations in a thesaurus can be divided into intrathesaurus relations within a coherent terminological system and interthesaurus relations between independent terminological systems. They are used to represent relationships between concepts and between concepts and terms, i.e. from the class HierarchyTerm to the class HierarchyTerm or Term.

The intrathesaurus relations identified by ISO2788 are: the *hierarchical relationships,* distinguishing a systematic thesaurus from an unstructured list of terms (glossary or dictionary), associating concepts bearing broader/narrower meanings, identified by the *BT(broader term)* relation, the *associative relationships*, relating concepts that are not members of an equivalence set nor can they be organized in a hierarchy, identified by the *RT (related term)* relation, and finally the *equivalence relationship* established between preferred and non-preferred terms, considered to refer to the same concept, and it is identified by the *use* and its inverse *UF* (used for) relations. As it does not distinguish between terms and concept, and we reinterpret these relations as the link between the conceptual and linguistic level. We refer in the following their functional role and specializations.

BT is used for semantic generalization or specialization of query terms. From a knowledge representation approach, the *BT* relation carries isA semantics, and a query term may be expanded by its narrower terms, if we ask for objects of this kind. Consequently SIS-TMS enforces that all HierarchyTerms have a broader Term except for TopTerms, and that the BT relation is acyclic. A Term may have multiple broader terms in the sense of multiple supersets. Thesaurus maintainers may distinguish between the main and alternate broader terms.

RT is used for the detection of relevant concepts by users. It plays a role like a general attribute category in KR systems. Dozens of useful specializations can be found, as the "subdivisions" of ISO2709, whole-part relations, and rule-related relations. In the latter case, machine interpretation may occur. *Art & Architecture Thesaurus* team has identified more that 20 different meanings of the RT relation.

UF (use for) can be used by users as entry points in a thesaurus. Actually most thesauri distinguish the *ALT* relation to preferred terms from the *UF* to non-preferred terms. The EET (European Education Thesaurus) [33] consequently regards any translation of a concept to some language as a kind of UF.

In SIS-TMS, *hierarchical association, equivalence association* and *associative relation* are modeled as metacategories of intrathesaurus links. These generic categories group and control the specialization of relations to preserve compatibility and to maintain the related global consistency rules. I.e. the application code can refer to those for constraint enforcement and for export of data in a compatible format. Hence the application code is robust against extensions. User defined extensions as the above mentioned specializations of *BT*, *RT*, and *UF* links are substantial for specific applications and the maintenance of their logical consistency, as well as for

the conceptual evolution of thesaurus structures into knowledge bases despite format standardization.

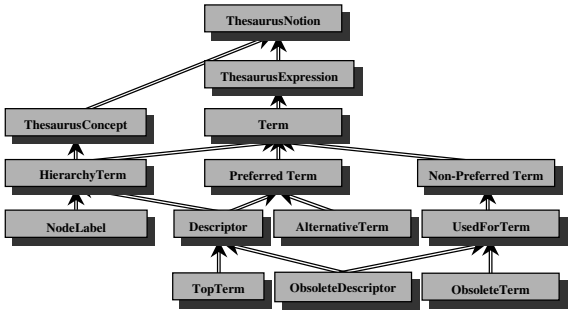


Fig. 2. IsA hierarchy of the SIS-TMS Classes of thesaurus notions

Other systems allow for new user defined links, but do not relate them to existing semantics and hence do not automatically imply standard conformant viewing, handling and constraint enforcement. Specialization of relations is a distinct feature of the SIS-TMS.

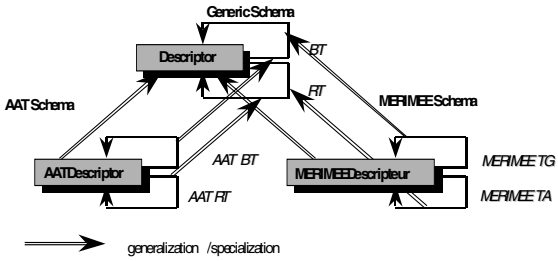


Fig. 3. Subschema Creation per Thesaurus

Representing Multiple Interlinked Thesauri. A challenge in the SIS-TMS was the development of the conceptual model to incorporate multiple interlinked thesauri under one database schema. For the development of the currently implemented conceptual schema two approaches were studied. The problems is, how to distinguish between the work of each thesaurus editor on one side, on the other side to see the common things as common without running continuously search routines, and on the third side to get a global view on how the different thesauri fit together. For more details see [11].

According to the second, a global name space is made up for all terms and concepts of one language. For each thesaurus, a separate schema is generated. Due to multiple instantiation in SIS, these schemata can overlap conflict-free on the data. The system tables stay limited per thesaurus. Gradual merge is possible without duplicating the records, as the same term can participate in any thesaurus. (Terms are not regarded as the invention of the thesaurus editor). Each thesaurus may have different semantic structures. Links are not confused, as they belong to individual schemata. Terms of different languages are distinguished by prefixes. Concepts of an interlingua can be dealt likewise.

On top of this world, one generic schema is developed as superclasses to provide global views. "Supercategories", abstractions of links and attributes from the individual schema as presented in the previous chapter, provide the notion of global semantics. For each new thesaurus, the generic schema is duplicated by specialization of its generic classes and relations into thesaurus specific ones, and the thesaurus data are loaded under these. For deviations in semantics, appropriate extensions can be made. This model results in a large schema. As an SIS schema is declarative, and not by storage allocation, no space is wasted. We selected this novel approach.

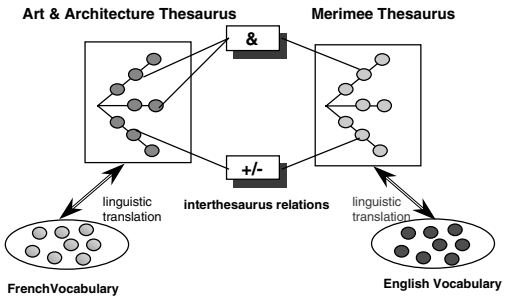


Fig. 4. Multilingual relations

As a consequence, we may see on one common descriptor all links made by other thesauri, achieving a kind of trivial merge. Each schema provides an isolated view of one thesaurus, and the generic schema a unified view of all. Basically, each thesaurus is handled as an annotation, an *opinion* of a group on a common term and concept space. This method can be elaborated into much more sophistication [11].

Interthesaurus Relations. The interthesaurus relations modeled in the SIS-TMS are an extended version of the ISO5964 links as presented in [10]. Those refined relations are the outcome of a discussion that took place in the framework of the AQUARELLE project between experts in multilingual thesaurus creation and ICS-FORTH, the provider of the multilingual thesaurus management component used in the project. The specific problem was to embed a system of multiple independent thesauri in different languages into a system for access to heterogeneous databases containing objects of material culture, supporting automatic term expansion/trans-

lation under a Z39.50 protocol. The conclusion of this discussion was that ISO5964 does not define precise enough semantics for that purpose.

We define the following relations: *exact equivalence, broader equivalence, narrower equivalence, inexact equivalence, union* and *intersection* of concepts. A detailed presentation of the semantics of above relations are presented in [10]. These links are from concept to concept (HierarchyTerm), and should not confused with linguistic translations, which use any suitable word from the other language rather than the specific thesaurus descriptors.

Obviously equivalence relations are opinions of one group, or at least under the responsibility of one group. Of course, good teams seek advice from each other. But the geographical distance and other local needs hinder synchronous updates. We therefore foresee different equivalence relations for group A from Thesaurus A to B, than for group B from B to A. If group A or B withdraws a concept, it remains marked as obsolete in the database, giving the other group a chance to redirect their links later. New concepts are marked as new, and should not be referred to until released. Suitable permissions can be set up in the SIS-TMS, so that such a database can be maintained cooperatively without conflicts through the net.

3.3 Maintaining Consistency between CMS and the TMS

In the environment presented in section 1 we foresee, that Collection Management Systems (CMS) such as digital libraries, library systems, museum documentation systems and others draw their classifation vocabulary from a Local Thesaurus Management Systems (TMS) which is a shared resource. The contents of the Local Thesaurus Management Systems have been initialized from the contents of the local CMS. The CMS needs continuously new terms, and incorporates classification terms in its records, eventually in central lists as well. The CMS can also propose new terms to the TMS. The TMS will be updated with new terms from many sides, and old concepts and terms may be renamed, revised and reorganized.

The essential problem is to ensure and maintain consistency between the contents of the vocabularies in the underlying CMS and the contents of the Local Thesaurus Management Systems. It can be regarded as a heterogeneous database problem with vertical distribution, classified data records on one side, semantic term structures on the other, and the shared identifiers both sides communicate on are the descriptors. Descriptors, i.e. concepts for classification, may be renamed, and the identity is lost. One could use system identifiers, numbers, instead, but this is the same impractical in distributed environments. Further, existing CMS do not necessarily have foreseen system identifiers for terms. Even the term codes the AAT uses are not preserved from version to version. The solution is simple: there are no global persistent identifiers. Instead, a history of identifiers is kept for each concept from release to release in the TMS. A CMS must note, with which release its terms agree. With this knowledge, concepts can permanently and automatically be identified between all systems in such a federation.

Further, abandoned concepts must be marked. In this case, an expert has to find in the records of the CMS which other concept applies in each case. Ideally, there should be links that indicate all shifts of the scope of concepts from version to version. Thesaurus editors are however not used to do so. At least all new descriptors mark an environment of concepts, where concepts may have changed scope or have been refined. As the respective updates in the CMS have to be done by hand, they are time consuming, and immediate consistency between CMS and the TMS is not possible. Suitable interface software cans speed-up the process of updating the CMS. The CMS does not need any knowledge of the semantic links in the TMS.

Finally, Term Servers or other TMS should to be kept up-to-date with data of a local TMS. Even though a thesaurus is a central resource, it changes slowly, and therefore it is practical to have local copies around. Term Servers, as described above, should combine different local authorities and in addition maintain interthesaurus links between those. Therefore, incremental updates have to be foreseen, even more, as authorities can be very large.

For each part of an authority, which is to be shared with another term base, only one management system must be the master. The master maintains as well the richest semantics, the others (slaves) may or may not keep reduced schemata. Nevertheless, the slaves may make references to the imported data. Therefore, as above, concepts cannot just be withdrawn. Beyond the above measures for communicating updates on concepts, all semantic descriptions have to be transferred. As those are attached in our system to the descriptors, it is sufficient for incremental updates to mark descriptors of the master which encountered a change in their attached information, and to transfer all the attached information to the slaves. Following these considerations, we have implemented the version control in the SIS-TMS.

Version Control and Data Consistency. The purpose of the version control is the information of thesaurus editors about previous discussions and states, and the capability to incrementally update another CMS or term server with the changes done in the TMS. Thesaurus releases are created at a slow rate, months or years. A rollback features is therefore not necessary, backups are sufficient for that purpose. Individual changes can be withdrawn at any time. A function is however provided, which inserts the latest changes into the last release for „last minute changes“.

Consequently, the idea of the SIS-TMS implementation is to keep in the database only the least versioning information for the above purposes as *backward differences* for scalability reasons. All other version data may be put in history logs in future versions. Versioning is based on releases rather than dates. The "current" release is being edited, and no history of changes is kept within it. Rather, the results of individual changes are merged. In contrary to version control systems, always the current version is displayed together with all registered backward changes per entity. The latter can be filtered out. Under this perspective, we register whether

- a descriptor has been introduced (a new concept is described)
- an existing descriptor has been abandoned (the concept is regarded inappropriate for classification or should be composed dynamically from other concepts)

– an existing descriptor is renamed
– any semantic information around a descriptor has changed.

We distinguish between operations on released concepts and on unreleased. In the unreleased, the user can introduce descriptors, which are classified as "*new descriptor*". He/she can perform all operations on descriptors and undo them.

The operations on released concepts are constraint, because they contain data that have been communicated to other systems and may have been used for indexing. Introduction of descriptors as well as deletions are not permitted. A descriptor can be abandoned by the following procedure: It is classified as *"obsolete descriptor"* and its *broader/narrower* associations to the others are deleted but it remains a member of the term list of its hierarchy, retaining the context in which it has been defined. Further, the „gap" in the hierarchie is „closed" by drawing *BT* relations between the narrower terms of the *obsolete* descriptor and its broader terms. We do not constrain any changes in the associated information, semantic and administrational links and attributes, as this is not necessary to keep other systems up-to-date or to avoid daggling references.

If editors regard the noun phrase, which identifies a descriptor, as inappropriate for the semantics of the concept, or it is going to cause name conflicts, it can be renamed. Renaming an object in SIS does not alter its system identifier and its properties remain attached to it. SIS-TMS maintains uniqueness of all term-names in the system. Whereas ALT and UF terms can be deleted at any time. Released descriptor names, which come out of use, become „obsolete terms". Further, for each release, links are maintained where the names have gone to. The complete algorithm is not trivial, as within one release all rename actions must be merged in order to be unique, and obsolete names may be reused as ALT or UF terms. Currently, following the AAT philosophy, we do not allow a concept (descriptor) to refer another descriptor as UF, but we expect in that case, that the respective descriptor is renamed to disambiguate it from the other concept.

3.4 Interaction in the SIS-TMS

The user interacts with the SIS-TMS via its graphical user interface, that provides *unconstrained navigation* within and between multiple interlinked thesauri. The user can retrieve information from the SIS-TMS knowledge base using a number of predefined, configurable queries and accept the results either in *textual or graphical form*.

The implemented predefined queries support access to the semantic and managerial relations of a term and to the contents of a hierarchy or facet. It has been considered by thesaurus users that the graphical presentation of the broader term relations is an essential requirement for thesaurus interfaces. The SIS-TMS not only provides such graphical representations but an essential feature is its ability to represent in a single graph any combination of relationships in arbitrary depth. (See fig 5, central window). A very efficient graph-layout algorithm allows the display of large structures in real time. A „global view" window is used to zoom in any part of a larger graph.

Therefore the user is able to view, in a single graph, all relations pertaining to a term. The existence of multiple broader terms introduce no difficulty to such graphical presentations. These graphs are of particular interest to authority providers since they represent in an compact way all the necessary relations for the detection of inconsistencies. The development group of the French MERIMEE thesaurus has already mentioned the easiness of identifying inconsistencies using the graphical representations of the SIS-TMS in contrast to scanning textual presentations.

The SIS-TMS graphical user interface does not need customization if a thesaurus is loaded in the knowledge base. The implementation of the predefined queries uses the generic schema presented previously and is therefore thesaurus independent . Consequently, if a thesaurus is loaded in the SIS-TMS knowledge base, it can be immediately queried without any customization of the predefined queries.

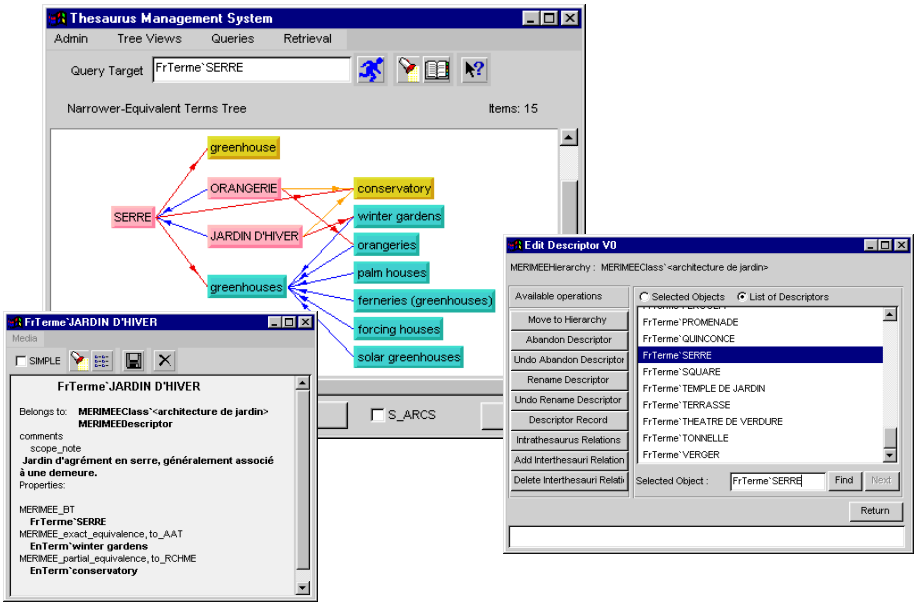


Fig. 5. SIS-TMS User Interface, Browser and Data Entry facility.

The loading of thesauri and term lists compliant with the ISO2788 and ISO5964 principles is performed by means of an input module. In the near future, we intend to develop a generic import tool that accepts as input thesauri described in other formalisms besides the above. The SIS-TMS has an output module that exports term lists. A flexible report writer that will be implemented in the near future.
The maintenance of thesaurus contents in the SIS-TMS is performed either with the import module supporting batch updates in the information base in form of tagged text files or interactive updates through the *Entry Forms*.

We distinguish *primitive* and *complex* operations: the former are the *creation, deletion, renaming, classification* (assignment to a class), *generalization* (assignment

to a superclass) and *attribute assignment* (creation of a link) and the latter are defined on the basis of more than one primitive operation.

The updates in the SIS-TMS are performed through the *Entry Forms* in a task-oriented way.(See fig 5, right window). A task is defined by a set of *objects*, which can be updated with this task, and a set of operations, which can be performed on these objects. Tasks can be configured at run-time. This approach results in a *logical partition* of the contents of the information base for user, which can be used to organize access permissions, in particular for the cooperative work on interlinked thesauri. The following groups of tasks have been set-up SIS-TMS:

- on the organizational part of a thesaurus
- on the contents of a thesaurus and
- on administrative information of a thesaurus.

Tasks concerning the organizational part of a thesaurus are performed on facets and hierarchies, which can be created, deleted, renamed as long as the defined consistency constraints are preserved. Contents tasks manipulate terms and semantic links and maintain the version control as described above. Administrative tasks concern the updates on the literary warrants (source references where a term is used) and editors responsible for changes of the thesaurus contents.

4 Experience and Conclusions

The SIS-TMS is a refinement of the VCS Prototype, a terminology management system developed by FORTH in cooperation with the Getty Information Institute in the framework of a feasibility study. Based on the evaluation of this prototype and new requirements, a new version of the system and the Semantic Index System server formed for the terminological system in the AQUARELLE system. In the sequence, installations have been made at thee French, Greek and Italian Ministry of Culture, the RCHME Thesaurus Group and the MDA. Further, the system has been used for access to digital libraries and traditional library systems in Greece.

The experience at these installations confirms the need to take thesauri out of the local databases, and to install systems, which allow to keep the use of terms consistent with evolving authorities. All these organizations maintain a series of databases, either under separate local responsibility or under central administration, with which they need to communicate terminology. Our vision of a three level architecture fits very well there. The AQUARELLE access system demonstrated strikingly the need to provide translation capabilities between different thesauri, as we provide, and not only uniform multilingual thesauri in interlingua. Further the value of real-time flexible graphics was intensely appreciated by all users. It could be demonstrated, that certain structural properties and associated quality issues can only be controlled by graphics. The data entry system proved to be appropriate. Further enhancement is needed in the report writing: users like to print out thesauri in a book-like form. More experimental results will be available at the end of the AQUARELLE project.

We have the impression that integrated terminology services in distributed digital collections are going to become an important subject, and that the SIS-TMS provides

a valuable contribution to that. It solves a major problem, the consistent maintenance of the necessarily central terminological resources between semiautonomous systems. The terminological bases themselves need not be internally distributed, as the access needs low bandwidth, read-only copies can easily be sent around at the given low update rates, and term servers can be cascaded. In the near future, we shall further enhance the functionality of this system to make its usability as wide as possible. In parallel, we engage and are looking for projects that integrate terminology services. Whereas there exist several standards for thesaurus contents, no one has so far tried to standardize the three component interfaces: (1) Term Server to retrieval tools, (2) TMS to CMS, (3) TMS to Term Server. As in a distributed information system many components from many providers exist, we are convinced that these three interfaces must become open and standardized, to make a wide use reality.

5 References

1. Jaana Kristensen, Expanding end-user's query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, 29(6):733-744, 1993.
2. A. Spink, Term Relevance Feedback and Query Expansion: Relation to Design. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 81-90, 1994.
3. Spink A., Goodrum A., Robins D., & Mei Mei Wu. „Elicitations during information retrieval: Implications for IR system design." In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 120-127). Konstanz:Hartung-Gorre, 1996.
4. Brajnik G., Mizzaro S., & Tasso C., „Evaluating user interfaces to information retrieval systems: A case study on user support." In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 128-136). Konstanz:Hartung-Gorre 1996.
5. P. Constantopoulos and M. Doerr, An Approach to Indexing Annotated Images, Multimedia Computing and Museums, Selected Papers from the *Third International Conference on Hypermedia and Interactivity in Museums*, by David Bearman, pp. 278-298, San Diego-CA, USA, October 1995.
6. A.F. Smeaton and I. Quigley. Experiments on using Semantic Distances between Words in Image Caption Retrieval. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 174-180, Zurich, August 1996.
7. D. J. Foskett. Thesaurus. In *Readings in Information Retrieval,* eds. K. Sparck Jones and P. Willet, publisher Morgan Kaufmann, 1997.
8. D. Soergel. „SemWeb: Proposal for an open, multifunctional, multilingual system for integrated access to knowledge about concepts and terminology." *Advances in Knowledge Organization,* 5, pp.165-173, 1996.
9. M. Doerr. Authority Services in Global Information Spaces. *Technical Report*, *ICS-FORTH/TR-163*, Institute of Computer Science-FORTH, 1996.
10. M. Doerr and I. Fundulaki. A proposal on extended interthesaurus links semantics. *Technical Report* ICS-FORT/TR-215, March 1998.

11. M. Doerr, "Reference Information Acquisition and Coordination", in: "ASIS'97 -Digital Collections: Implications for Users, Funders, Developers and Maintainers", *Proceedings of the 60th Annual Meeting of the American Society for Information Sciences*, " November 1-6 '97, Washington, Vol.34. Information Today Inc.: Medford, New Jersey, 1997. ISBN 1-57387-048-X.
12. V. Christophides, M. Doerr and I. Fundulaki. The Specialist seeks Expert Views-Managing Folders in the AQUARELLE project. Selected Paper from the *Museums and the Web, MW97*, eds. D. Bearman, J. Trant.
13. M. Doerr, I. Fundulaki „The Aquarelle Terminology Service", ERCIM News Number 33, April1998, p14-15
14. R. Kramer, R. Nikolai, C. Habeck. Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. In *International Journal on Digital Libraries* (1), pp. 122-131, 1997.
15. G. Wiederhold. Interoperation, mediation and ontologies. In Proceedings of the International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative Knowledge-Bases (ICOT), Japan, December 1994, W3, pp.33-48.
16. R. Niehoff and G. Mack. The Vocabulary Switching System. In *International Classification*, 12(1):2-6, 1985.
17. A. Stern and N. Richette. On the construction of a super thesaurus based on existing thesauri. In Tools for Knowledge Organisation and the Human Interface. Vol. 2, pp. 133-144, 1990.
18. H. H. Neville. Feasibility study of a scheme for reconciling thesauri covering a common subject. In *Journal of Documentation*, 26(4), pp. 313-336, 1970.
19. R. Rada. Connecting and evaluating thesauri: Issues and cases. *International Classification*, 14(2), pp. 63-69, 1987.
20. R. Rada. Maintaining thesauri and metathesauri. International Classification, 17(3), pp. 158-164, 1990.
21. C. Sneiderman and E. Bicknell. Computer-assisted dynamic integration of multiple medical thesauruses. *In Comp. Biol. Med.*, 22(1), pp.135-145. 1992.
22. M. Sintichakis and P. Constantopoulos, A Method for Monolingual Thesauri Merging, Proc. of the 20th International Conference on Research and Development in Information Retrieval, ACM SIGIR, July 1997,Philadelphia, PA, USA.
23. S. Bechhofer and C. Goble. *Art Position Paper* - The Need for Structured Terminology. Medical Informatics Group, University of Manchester, March 1997.
24. E. H. Johnson and P. A. Cohrane. A Hypertextual Intefrace for a Searher's Thesaurus, Grainger Engineering Library Information Center, University of Illinois, June 1995.
25. Getty Information Institute (1995) Request for Comment Issued for the New Vocabulary Coordination System for the Getty Information Institute Authorities. Santa Monica, CA. (http://www.gii.getty.edu/gii/newsarch.html#article6).
26. A. Michard, G. Pham-Dac, Descriptions of collections and encyclopaedias on the Web using XML. To be published in *Archives and Museums Informatics*, Kluwer Pub., 1998.
27. The AQUARELLE project, TELEMATICS Application Program of the European Commission, Project IE-2005 1996.
28. P. Constantopoulos and M. Doerr. The Semantic Index System - A brief presentation.
29. J. Mylopoulos, A. Borgida, M. Jarke, M. Koubarakis, Telos: Representing Knowledge about Information Systems, *ACM Transactions on Information Systems,* October 1990.

30. A. Analyti and P. Constantopoulos and N. Spyratos. "On the Definition of Semantic Networks Semantics", *Technical Report*, Institute of Computer Science-FORTH, ICS/TR-187, February 1997.
31. R. S. Michalski. Beyond Prototypes and Frames: The Two-Tiered Concept Representation. *Categories and Concepts, Theoretical Views and Inductive Data Analysis*, eds. I. Mechelen, J. Hampton, R. Michalski, P. Theuns, 1993
32. D. Soergel. The Arts and Architecture Thesaurus (AAT)-A critical appraisal. *Technical Report*, College of Library and Information Sciences, University of Meryland, 1995.
33. C. Roulin. Sub-Thesauri as part of a metathesaurus. *In International Study Conference on Classification Research, Classification Research for knowledge representation and organisation*, pp. 329-336. Elsevier, 1992.
34. Introduction to the Art & Architecture Thesaurus. Published on behalf of The Getty Art History Information Program, Oxford University Press, New York, 1994.