A Metastructure for Thesauri in Archaeology

Martin Doerr

Institute of Computer Science, Foundation for Research and Technology – Hellas

Heraklion, Crete, Greece

e-mail: martin@ics.forth.gr

Demetrios Kalomoirakis

Greek Ministry of Culture, Athens, Greece

e-mail: demetrios.kalomoirakis@damd.culture.gr

Abstract

Controlled terminology is becoming increasingly important for archaeological documentation and information retrieval in wide area networks. Even though there is plenty of literature on thesaurus creation, the logical organisation, the "metastructure" of a thesaurus is still rather unreflected, in particular in the more abstract layers. So-called facet indicators are distributed intuitively in order to group a set of sibling terms by an implicit property. They can be regarded as potential elements of a metastructure. In co-operation with ICS-FORTH the Greek Ministry of Culture is developing a general-purpose thesaurus for archaeology. An extensible, systematic metastructure for terminology in archaeology is proposed, based on contextual and phenomenal notions. This seems to be more comprehensive than traditional approaches. This paper discusses the benefits of such an approach. In particular, the proposed structure seems to provide a better understanding of the implicit criteria intrinsic to object classification terms and the relation between compound expressions and traditional terms.

Key words: archaeological terminology, facetted classification, thesauri, qualia, subsumption hierarchies, concept formation

1. Introduction

Controlled terminology is becoming increasingly important for electronic documentation of archaeological object description and its consistent retrieval in wide area networks. In particular thesauri, which organises terms or concepts in semantic networks, is becoming a mandatory tool for users to find their way through the rapidly growing electronic information flood and for unambiguous communication through automated intermediaries. Even though there is plenty of literature and an established practice on creating thesauri (see e.g. publications of the Getty Foundation, the British Arts and Humanities Data Service AHDS, and numerous national standards), their logical organisation, the metastructure of a thesaurus is still rather unreflected, in particular in the more abstract layers. In this paper, we will try to clarify some concepts and present the logical organisation of a general purpose thesaurus for archaeology developed at the Greek Ministry of Culture in co-operation with ICS-FORTH. Finally we will point out some research topics, that came up with this work.

2. What is a Thesaurus

The notion of a linguistic thesaurus has been introduced over one hundred years ago by Peter Mark Roget. In 1852, he published his famous "Thesaurus of English Words and Phrases", intended to help authors find synonyms or more precise terms. Such thesauri are ever since quite popular, and Roget's one is still being edited [Encarta].

With the raise of electronic communication, and in particular with the search for information in database records, the need to refer to some concepts with standardised expressions also arose, otherwise entries made by curators and requests entered by users would never match. The currently most successful approach to standardise expressions are the so-called thesauri in the sense of computer science.

At this moment two directions can be distinguished: a linguistically oriented and a conceptually oriented. The "linguistic" one regards the "term" as the key element of the thesaurus, coming from a dictionary world. Terms are represented as the expert would use it in speech: potsherd, vessel, column, pottery etc. Besides textual descriptions of the term meaning, linguistic information is also added, such as part-of-speech, sort keys etc.

Respective data standards have been promoted by the Text Encoding Initiative [TEI], the latest developments being ISO 12620, MAchine Readable Terminology Interchange Format (MARTIF), and the Virtual HyperGlossary (VHG) (Murray 1998:34). Terms are typically organised in mono hierarchies (i.e. one term belongs exactly to one broader term) as a kind of associative decision trees leading from higher to lower abstraction levels of conceptually close terms. These hierarchies serve mainly as searching aids for the user to find a term rather than expressing abstraction levels. Often information is added to create a nice lay-out of indented lists in printed reports.

These thesauri might be helpful for individuals looking up and understanding terminology as an encyclopaedia, for linguistic work, translations and natural language search. However, there are severe drawbacks for indexing and classification in databases. The naïve assumption that a term uttered by an expert is a sufficient surrogate for the item we are describing or looking for does not hold true from the following reasons: terms are context dependent, be it true homonimity ("contrastive ambiguity" (Pustejovsky 1995:27)), shift of meaning ("complementary polysemy"), or because the term expresses not all aspects of an item relevant for the one or other future user. A "vessel" can be a number of different things, "pink" is not always a colour etc. A "temple" from the Greco-Roman period may have different associations and abstractions from a Protestant one. “Neo-classical building” and “school house” can be equally relevant characteristics of the same object. Further, users often query at a higher level of abstraction than the most detailed classification of an object, e.g. “black vases from Attica of the third and fourth century BC”. Hierarchies that are not simply associative, but that are built on true abstraction, can help the software to resolve such queries automatically through the thesaurus.

Such considerations lead the middle of the twentieth century library scientists to the development of classification languages based on concepts rather than terms. As C. Welty (Welty and Jenkins 1999:155) pointed out, only since shelving books on unique places is no more the major organisation principle, assignment of multiple subjects/terms were considered. So on one hand, the problem was addressed by the rules to combine “elemental” concepts either already in the terminological system (pre co-ordination) or at time of classification (post co-ordination). For this purpose the classification vocabulary was grouped into so-called “facets”, i.e. sets of elemental concepts, going back to Ranganathan’s Colon Classification System (Ranganathan 1965). On the other hand, these “facets” were internally organised by conceptual relations. Such systems form *thesauri* in the sense we will use from here on. As Pollitt (Pollitt 1997a) stated:

“Problems in the Verbal Plane brought about the introduction of regularisation and structure into the language used in post coordinated systems via thesauri.” Vickery (Vickery 1960:181) reports how he first encountered the word thesaurus:

> “Speaking at the Dorking conference on classification, Helen Brownson said ‘The problem (of information retrieval), as some investigators see it at least, is to transform concepts and their relationships, as expressed in the language of documents, into a somewhat more regularised language, with synonyms controlled and syntactic structures simplified. Now it is reasonable to think that the further we can go in routinising and mechanising the techniques of translating ordinary language into a regularised language and of coding for machine manipulation, the more we will be likely to achieve economically feasible machine searching on a large scale…[Some investigators] have come up with the thought that the best answer … may be the application of a mechanised thesaurus based on networks of related meanings.’ In the same paper Vickery (1960:185) cites Bernier and Heumann (1957) as proposing the introduction of an organised vocabulary in the form of a thesaurus.”

This “conceptual direction” of thesauri, typically elaborating on the semantic links as defined in ISO2788, ISO2709 and ISO5964, regards the “concept” as a key element of the thesaurus (see e.g. Svenonius 1989:82). These concepts can be alluded to by a series of linguistic expressions, but the concept itself has no name. In order to improve the orientation of the user, one term may be selected arbitrarily as an *identifier* for each concept (often called “*descriptor*”), eventually giving the impression as if the thesaurus were still about the term. This idea is also well explained in the explanations to the Art and Architecture Thesaurus AAT (Getty AHIP 1994), whereas ISO2788 still has no clear notion of separating between terms and concepts. WordNet (Miller 1993) identifies the concept (here called “meaning”) by the combination of *all* appropriate terms. Miller et al. make clear the excessive relations between words or terms and concepts.

Finally it should be mentioned, that there is an automatic generation of the so-called “thesauri” (e.g. Hazewinkel 1996, Chen 1993:25). By statistical means, measures for a conceptual distance between terms or expressions are derived based on the co-occurrence *in corpora* and by grammatical criteria. These methods do not give much insight into the semantic nature of the derived relations, and should be regarded more as retrieval aids than thesauri. Nevertheless, in the future such methods may be quite useful in the creation of a true thesauri.

3. About concepts

Cognitive psychology scientists have proposed several definitions for the notion of “concept” (see also Doerr 1998). According to the first point of view, a concept is a person’s conceptualisation of the notion of categories (the class or set of entities which are grouped together on the basis of some criterion or rule (Michalski 1993). According to this point of view, there would have to be as many concepts for the same category as there are different cognitive representations for it. For thesauri we do not adopt this approach since the thesauri incorporated are collections of terms with generally agreed on semantics and not individual definitions or representations of a concept.

According to the second point of view, a concept is perceived as a set of entities, called “concept instances” by a common agreement rather than formal reasoning on the properties which may make an entity the instance of the concept. In other words, we regard a concept as defined within a group of people, as long as the people agree which items are characterised by the concept. This does not exclude the use of rules, however it is not based on it. It differs to the first point of view in the fact, that we do not argue about what people really conceptualise, but what they can agree to on the outside. This is actually the basis for a semantic thesaurus construction. It is further a principle, which can be easily applied to the classification of material objects. Eventually we are forced to communicate such concepts to each other by a verbose account of examples and contextual information, called (in the context of thesauri) “scope notes”.

According to the third point of view, a concept is identified by a classification rule. This means, that the concept is further analysed into a logical expression as regards its properties. However, quite often conscious rules cannot easily be formulated even for obvious concepts, e.g. the optical characteristics of an aquarelle, and the rules themselves cannot be further analysed. Nevertheless, this principle is the basis for the so-called “Terminological Logic” or “Description Logic”(Borgida 1995:671), [DL], which can be regarded as the most advanced extension of conceptual thesauri. It has been proven quite successfully in medical applications (Rector 1997) as well as other domains. In our opinion it is more appropriate for cases where distinct overt features dominate, like in scientific objects and technical artefacts. As I will show below, it is quite useful to complement an agreement-based body of concepts by the more dynamic rule-based expressions.

Advantages of the conceptual approach are, that the conflicts of homonymity, as well as cultural and contextual differences can be overcome or made explicit. What counts is the definition of the concept rather than the term itself. Consequently, the AAT team has created expressions like “pink (colour)”, which are not natural, but resolve ambiguity. Further, we can identify clear abstraction hierarchies based on semantic inclusion (generalisation, super-class, subsumption, hyponomy and whatever synonym term may be in use), that allow to expand more abstract database queries into sets of specific terms. E.g. “dime” is a “coin”, “baby doll” is a “recreational artefact”, “kachina” is a “figural work” etc.

3.1. About objects and subjects

As we can see from the historical considerations above, library science, and in particular the problem of subject classification, has been the driving force behind these developments. Museums and archaeology, dealing to a very large extent with material objects, have relatively unreflectedly adopted the same method to classify their objects instead of the literature about them. At this point let us make a few propositions as regards the differences:

- Physical objects are not “abstract”. An “Introduction to Biology” and “The Behaviour of Geese” can be the distinct subjects of two different books. In case of objects, there is no single “living being” e.g., without being something concrete, like “goose”.
- Consequently, a physical object is truly classified by any higher abstraction of its type, but only the lowest has instances. My dog is a dog, a carnivore, an animal, a living being, a material object. Any existing animal however must be of the most specific type, whereas a book about animals may be as generic as the concept “animal”.
- The lowest abstraction level is relative, a function of our attention to more or less characteristics. It depends on the size and constitution of the collection at our disposal. The relatively well-defined level of a “species” in biology has only weak analogies in archaeological objects.
- An object may be classified by an open number of relevant views, according to the investigators objectives. These views may be completely uncorrelated as e.g. “exchange item”, “caouri shell”, and “property of an Oba”, “bead”, “element of a preserved string”, “slightly worn”. The “subject” on the other side may be seen as *one* such view, valid for literature objects, but mostly absent in material objects.
- Even though any object can give raise to a subject in the sense of “talking about…” (Welty and Jenkins 1999:155), the “parallel hierarchy” (Soergel 1995:369) generated by objects viewed as subjects may differ. E.g. talking about “bridge construction” is a kind of talking about “bridges”. “Bridge construction” itself is not a kind of “bridges”, but a kind of construction, either a design or an activity or both.

Of course, any piece of literature as a whole has its own unique individuality similar to a museum object (Welty and Ide 1999). We however agree here with (Welty and Jenkins 1999:155), who prefers to see the subject of a certain book not as a parallel individual associated with each book, but as a collective concept applicable to more than one book. He further points out that in libraries fiction itself is mostly not classified with subjects. The focus of libraries and museums is obviously quite different, and give raise to endless misunderstandings between both communities.

To summarise, we are under the impression, that conceptual hierarchies for material objects can be and need to be stricter defined with respect to the notion of abstraction (hyponomy, broader/narrower terms) than usual in library subject catalogues, and that they need more structure reflecting different aspects of investigation. E.g. we cannot declare, “Top: Arts: Classical Studies: Journals” and “Top: Arts: Classical Studies: Academic Departments” as narrower terms of “Top: Arts: Classical Studies” (see the DMOZ project [DMOZ]), or regard bridge construction as specialisation of bridges (LCSH), but must separate entities of the real world by their deeply distinct nature.

3.2. Descriptive and characteristic concepts

The examples above suggest, that we may be able to separate concept hierarchies by the aspects they express, about people for instance by sex, race, intellectual properties, etc. This can be easily performed with concepts, which represent more or less verifiable properties of the one or the other kind, which *describe* basically an object by observational criteria imposed by the researcher. See e.g. the definition of “knives (weapons)” in the AAT: “Weapons designed primarily for cutting, consisting basically of a single-edged, pointed blade and a grip mounted asymmetrically in relation to the axis of the blade, closer to the back edge”. The concept of “weapon” itself is purely functional, and bare of any assumption about its physical form. A dynamic (post co-ordinated) term like “railway museum” expresses no more than what it says (see the AAT).

However, we also have to simultaneously operate with historical concepts of things *characteristic* for a kind of “design model” of a culture, like an “ushabti” or “kachina doll”. These concepts comprise of a series of aspects and implications. “Persian rugs”, e.g. exhibit the subtlety that not all rugs from Persia are Persian Rugs, nor are all Persian Rugs from Persia, even though at least a “genuine” one should be. Obviously the provenance “Persian” does not imply that a rug necessarily has the characteristic design observed in Persia, and the characteristic design may be copied or truly followed outside of Persia. The latter has a history, evolution and social context.

In this sequence let us talk about *descriptive* object concepts, as those, that analyse our observations about the objects, that have come upon us from the past, be it on features that can be at present verified on the item itself, or features that originate in historical knowledge.

In contrast, let us talk about *characteristic* object concepts, as those, that designate a “design model” of a culture, which has specific functions, designs and contexts of use altogether, has characteristic traditional names (e.g. “alcazar”), and varies over regions and evolves over times. The “ushabty” e.g. was initially a surrogate for the mummy, and later became a servant for the after-life. It has a characteristic span of forms and sizes. “Ushabty” (ushebti, ushabti, shawabty etc.) are transcriptions from the original Egyptian terms for that concept. In other cases, like with Neolithic objects, those are lost, and researchers “coin” modern terms as surrogates.

As we see, our object concepts come from different perspectives of cognition, with more complex relations between each other and the objects they classify, than simple hierarchies can appropriately capture. The results of pressing those into a simple hierarchy are a disorientation of the user trying to find concepts and poor or even questionable conclusions from the given semantic relations.

Ranganathan (Ranganathan 1965:38), characteristically states:

> "…that the work gets resolved into three different planes: the Idea, the Notional, and the Verbal Planes. Looking back from this position into the work done before these three planes of work were clearly seen and separated, it is found that much of the difficulty arose out of frequent, listless, and unconscious change from one plane to another. Another cause of difficulty was the inhibition of work in the Idea Plane by the limitations of the Notational plane, and by the inherent defects of the Verbal Plane." (see also Pollitt 1997a).

In this paper, we want to investigate some properties of the Idea Plane, the world of concepts and meanings, and their relation to objects of our investigation, free of limitations of the Notational plane, and in sequence propose some improvement in the Notational plane.

4. About facets

We have already mentioned the so-called "faceted classification". The term "facet" alludes to the fragmented surface of a diamond, suggesting that the same thing can be seen from different sides. Faceted classification and further developments in the form of Description Logic are becoming increasingly popular in the dealing with the complexity of precise capturing of information contents (e.g. Prieto-Diaz 1987:6, Constantopoulos 1995, Welty and Ide 1999 and Pollitt 1997a). Actually, one may distinguish three slightly different notions of facets with at least two different applications.

Following (Prieto-Diaz 1987:6): "The faceted method… relies… on… building up or synthesising from the subject statements of particular documents. By this method, subject statements are analysed into their component elemental classes, and these classes are listed in the schedule. Their generic relationships are the only relationships displayed."

This is the first notion of facets: *Elements for synthesising* complex (post co-ordinated) terms or statements by enumeration. In our opinion, this kind of analysis results more in grammatical elements, such as subject, predicate, object, etc. (see e.g. Constantopoulos 1995), than in "elemental classes". They become elements of an indexing language used to create an open number of potential "compound terms". Typical library facets are: Topics, periods, places, genre.

Whereas statements often summarise a scientific paper quite well, material objects are better characterised by noun phrases. Soergel (Soergel 1995:369) discusses in much detail the use and limitations of an indexing language (the AAT), which foresees only one generic relation between each facet. An example for such a compound is "nineteen century Massachusetts wood chairs". Please note, that this denotes no more than a chair. The AAT distinguishes the facets: "Object, Agents, Activities, Styles and Periods, Materials, Physical Attributes, Associated Concepts". Description Logic is the natural extension of such indexing languages, as it enriches the set of relationships allowable between the concepts and puts them in a well-defined logical framework (see e.g Welty and Jenkins 1999 and Bechhofer 1999 as regards the utility of DL).

The separation of our concepts into *disjoint cognitive categories* gives raise to a slightly different interpretation of "facet". Under this aspect, we regard as facets notions of objects, actors, events, measures, time, space, etc. "Facets" are seen as the building elements of our conceptualisations of a domain, without an immediate syntactic purpose. Such elemental classes are often referred to as "*major facets*". Ranganathan talks about "the five fundamental categories PMEST (Personality, Matter, Energy, Space, Time)", and the AAT facets seem to be of that kind. The difference as regards the above interpretation becomes obvious, when someone regards grammatical subjects and objects, which are identical in nature, but different in their role within a statement. E.g. a physical object can be the subject and the object in a grammatically correct phrase.

For an archaeological thesaurus, we regard the principle of "major facets" as appropriate to bring the first order into all potential concepts of the domain. The ontology of the CIDOC Conceptual Reference Model (Doerr 1999) is an attempt to provide a formal standard model used to describe objects and their history, mainly to be able to mediate and transform between different data and metadata formats. It distinguishes: *Temporal Entities* (including periods, events and activities), *Actors, Physical Objects, Conceptual Objects, Place* and *Time*. Recently the IFLA and the Dublin Core community came up with almost identical categories (see: the Indecs project, http://www.indecs.org/results/model.htm). The obvious similarity encourages standardisation. As Steven Pinker (Pinker 1994) suggests innate categories of human perception are most probably hidden behind this. We would however not dare say, to which degree nature itself dictates those. It is easier to derive facets for synthesising complex terms for different applications from these categories, than from purely syntactically motivated categories.

There is however yet another interpretation of "facet", which is, in our opinion completely different. Prieto-Diaz (Prieto-Diaz 1987:6) continues in the same paper: "Facets are sometimes considered as perspectives, viewpoints, or dimensions of a particular domain", and continues with an example about animals classified by habitat versus genealogy. Actually this is the notion that Ranganathan originally introduced, which indeed reminds us of facets of a jewel. The MDA Archaeological thesaurus [MDA97] e.g. introduces a term "armour by construction", and below "scale armour", etc, "armour by form", and below "cuirass", etc, "armour by function" and below "parade armour". Such terms, that announce the criterion by which the subsequent analysis is carried out, are called node-labels (ISO2788), *minor facets* or guide-terms (AAT). Frequently, the values these criteria can take on are taken from another "major facet". As e.g. in the term "cutting sword", the value "cutting" appears in the major facet "Activities" of the AAT.

4.1. The minor facets in the AAT

Ranganathan observed, that an open number of concept arrays can be used to provide the criteria for a facet on another array. Similarly, the AAT has systematically introduced guide terms (facet indicators) in a great number, to be precise 2840 in the 1998 edition. From a scientific and philosophical point of view, this situation is not very satisfactory. In the first statistical investigation, we have identified that there are a number of predominant criteria, that can be used to create a metastructure.

For that purpose, we have separated the guide terms of the object facet. We separated approximately 1640 of them (some concepts are not really objects, such as "illustrations", "museums"). From these, 615 actually refer to the criterion with the characteristic connector "by" like "swords by function". These could easily be separated, and became the object of our analysis. A more thorough analysis could well be done on the total, but at this stage we are only interested in a qualitative result:

318 guide terms out of the 615 fall in three clearly distinct dominant groups:

…by form:	129
…by function:	121
…by location or context:	68.

The remaining have frequencies below 15. The notion "location or context" is not easy to interpret. It is often a geometric relation of a part to a kind of whole, but also of independent smaller elements to a larger structure such as landscape to sea, lamps to parts of a room, etc. The other categories are clearly defined. In practice, form and function can overlap in the form of "functional forms". There are 15 guide terms "…by form or function", like "furniture by form or function" and others.

This analysis is only linguistic. Actually a closer look on the actual narrower terms appearing under the less frequent kinds of guide terms reveals more of the major categories, such as "…by shape", "…by merchandising practice", "…by position". The AAT has never tried to standardise these terms. Further, a large multitude of differently named categories can be characterised as construction criteria, be it the internal structure or the process of creation of the internal operation principles of machinery. We have not tried to go into higher detail, nor have we tried to distinguish event-like functions like "cutting" from activity-like functions like "dwelling". Construction and form sometimes overlap, in the case of technically necessary forms of machinery we have preferred to classify the criterion as constructive.

We have arbitrarily added the group "form or function" to "function". There is a small group of criteria about the social context of creation or use. Surprising enough, only 5 times the "subject type" appears in the proper sense, e.g. for sculpture, as a term-generating criterion! For example we have classified the "subject" of a museum under "function". Here are the statistics of our interpretation of the AAT object guide terms with the "by" connector:

Criteria of form:	34 %
Criteria of function:	29 %
Criteria of relative location:	15 %
Criteria of construction:	15 %
Criteria of social context:	5 %
Criteria of subject type:	1 %
Criteria of naming (like coins):	1 %.

From the impression these statistics give we derive the hypothesis, that there are a few predominant aspects to analyse objects, and not an arbitrary, unpredictable mass. These aspects are reflected in the terms themselves, but may be as well explicit criteria (distinct data base fields) of classification (function, form, material etc). Some British museums have made the functional aspect explicit: they classify objects with SHIC, (Social, Historical and Industrial Classification) [SHIC93] only by their relevance to human activities - a clear indication of the relevance and utility of the "function" and "social context" aspects. In this practice, the actual "function" is always explicit, and not hidden in some traditional term.

The above analysis gets an unexpected confirmation from modern linguistics. Pustejovsky (Pustejovsky 1995) has developed a theory to cater for the dynamic change of word meaning in new contexts. He shows, that nominals in general, and in all languages imply multiple meanings, which are "activated" in a sentence by the appropriate context, e.g. a verb that refers to a certain aspect. An example he gives is:

> "He walks through the door" (function, opening), and "He paints the door" (object).

He talks about the "Qualia Structure" of nominals, which he analyses in the following categories (referring also to Aristotle's notion of modes of explanation):

- Constitutive: The relation between an object and its constituents. (Material, weight, parts and component elements)
- Formal: That which distinguishes the object within a larger domain. (Orientation, magnitude, shape, dimensionality, colour, position)
- Telic: Purpose and Function of the object. (The purpose the agent has in performing an act. Built-in function or aim which specifies certain activities)
- Agentive: Factors involved in the origin or "bringing about" of an object. (Creator, artefact, natural kind, causal chain)

The similarity is striking. If eventually, common sense in understanding language and classification in archaeology will lead to the same or at least similar categories, will be shown in future research. By sure, both disciplines can learn from each other.

5. A metamodel for object thesauri

The fact that traditional terms silently imply one or the other aspect can bring confusion into the classification. Some data about an object may be classified under the above aspects with descriptive concepts, especially if a characteristic term for it is missing. Another object may be classified with traditional characteristic terms that imply a descriptive concept shared with another object. We shall not be able to retrieve both under the same terms.

For example, let's look at the AAT Terms "armor", "parade armor" and "garnitures". The scope note of "garnitures" says: "Sets of armor, …for use in warfare or tournaments. They were invariably made for important persons according to a precise design for specific types of events." Neither the aspect of protection in combat nor the aspect of social representation are explicit from the wording or from the broader terms given. On the other hand, a "pure descriptive" classification with the above criteria is neither efficient nor always appropriate. Archaeologists at the Greek National Archive of Monuments and others [PrivCom] have repeatedly referred to us with the problem, that the characteristic terms imply various properties not accessible in a database search. On the other hand, explicit reference at each instance is not economic, and implicit properties may change with the period even for the same term.

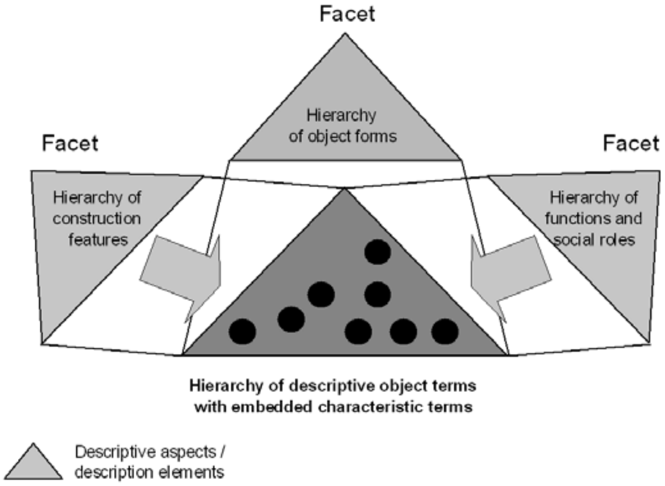
Facet
Hierarchy of object forms
Facet
Facet
Hierarchy of construction features
Hierarchy of functions and social roles
Hierarchy of descriptive object terms with embedded characteristic terms
Descriptive aspects / description elements

Figure 1: Conceptual Scheme of a multi faceted Hierarchy.

Under this consideration we propose to systematically classify characteristic terms by descriptive terms. With this practice, direct classification of objects with descriptive terms and classification with the correct characteristic term can become consistent and compatible. As we have seen above, a thesaurus can serve as a language indexing device and as an information source as regards the conceptualisation of a domain. For the latter, systematic classification of descriptive terms by all aspects (qualia) is a remarkable finding aid for users, as detailed in section 6.

An object clearly identified as a model of its culture, like an ushebti, is of course best characterised by its vernacular, historical term. Only that will render all relevant aspects. An object, which cannot clearly be identified as one or the other, should better be classified by all possible aspects, in order to provide future researchers with all potential candidates for their requests (ensure recall). Similarly, a term with multiple potential meanings should be generalised (i.e. assigned broader term relations) by all possible meanings. Equally prominent in archaeology are cases, were we try to identify unknown models of a culture, in particular of prehistoric ones. In such an event, we may use combinations of descriptive terms and eventually invent a new characteristic term, a "coined term", when our knowledge about the kind of items seems to stabilise.

This is not in contradiction to defining its proper scientifically established meaning by respective scope notes. This double practice not only helps to explain vernacular terms or their equivalents to the interested scholar, but also provides the correct generalisations for query expansion in the use as an indexing language. Even more, shifts in meaning could be expressed by specialising a term into its local and temporal variants, and generalising them differently. For example the most ancient Egyptian items that we have encountered are burial items, but their later equivalents are not [PrivCom], (contemporary secular equivalents are typically not preserved).

5.1. Examples

In order to illustrate this theory, let us in the following assume, that the relevant qualia are: function, form, construction. To create the proposed metastructure, we may provide an initial vocabulary of expressions of function (like SHIC), of forms, and of constructive features relevant to the set of objects under consideration. From these, we create the set of classes of objects, which are characterised by precisely one feature of the initial vocabulary, e.g. "objects for warfare", "objects for eating and drinking", "rectangular objects", "built objects". If the initial vocabularies are organised by hyponymy (broader/narrower terms), this procedure creates a "parallel hierarchy" (Soergel 1995:369) for each aspect (quale). These would be the primitive descriptive concepts. Each hierarchy could be headed by a node label like "objects by function" etc., even though the practice is to enumerate only a flat list under each node label. Any characteristic object concept should be describable as a narrower term of some combination of these classes. In this theoretic picture, characteristic concepts appear dispersed in the "pool" of all possible combinations of the primitive descriptive concepts, as figure 1 suggests.

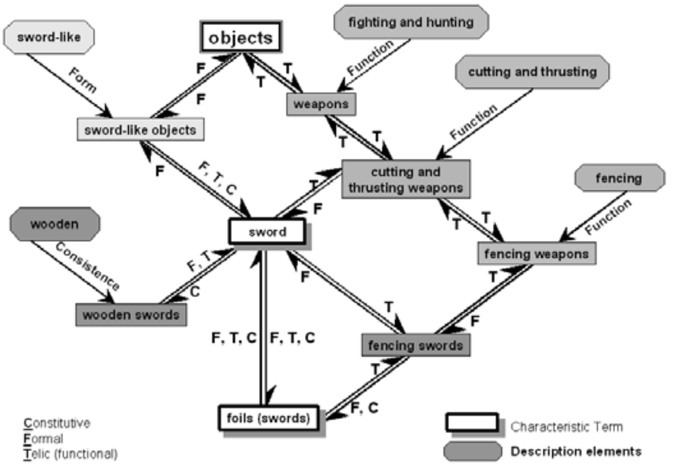

Figure 2: Sample hierarchy with typed broader/narrower relations.

In practice, two factors cause severe distortions to this picture: Properties of the formal aspect and the relations between the characteristic concepts. The formal aspect or quale seems to play a dominant role in the discrimination of characteristic concepts. Pustejovsky refers to it as "basic", without further justification. We observe a basic difficulty in the defining of expressive sets of shapes without the reference to objects types, however we easily derive shapes from object types: we may talk about a cross-shaped ground plan, a hammer-like fish ("hammerhead shark"), etc. This does not hold true for the other qualia. This is further confirmed by the dominance of form criteria in the AAT. As described in chapter 6, the Greek National Archive of Monuments has introduced the formal aspect as «ειδος», literally translated as "kind, model or article".

It is beyond the scope of this work to analyse the deeper reasons behind that. It seems economical to take characteristic concepts themselves per default as sufficient to represent their formal aspect. Consequently, an independent vocabulary of form is only created for those forms not captured by the characteristic terms themselves, and the total of characteristic terms is seen as a source to generate respective form expressions, e.g. with a "like" operator.

Beyond that we observe, that characteristic terms of some higher level of abstraction may be specialised themselves by one or the other aspect, or all aspects together. In the AAT, the guide terms appear characteristically at the top and then at any level without a recognisable system. Our interpretation is that actually any characteristic term is potentially specialised by all aspects at any level, however, these potential places in the polyhierarchical grid may not always be occupied with actual established terms. The current practice to introduce a node label in order to announce the aspect of making narrower terms (NT) has two limitations: firstly, it cannot be continued on the next level. Another node label is needed. Secondly, it does not say, which aspect of the inversion, the broader term relation (BT), entails. This is actually not of a trivial nature. There are combinations of all aspects between the two directions of a BT/NT relation. The broader term relation is actually the one that classifies the characteristic concept, whereas the narrower term relation is the one we need to systematically create the thesaurus top to down and to expand higher concepts into individual classification concepts.

Instead of node labels, it would be better for us to introduce different types of BT/NT relations, a real metastructure feature, which goes beyond ISO2788. The idea is illustrated in figure 2, around the characteristic concepts “swords” and “foils (swords)”. A type determinator in brackets accompanies the latter term, in order to disambiguate the concept from homonyms. The right branch (weapons…) shows part of an object hierarchy induced by activities “fighting and hunting”, “cutting and thrusting”. These pure functional concepts become specialised by the characteristic term “sword”, which adds a morphological and constitutive aspect. The latter is again specialised by the functional aspect “fencing”, and in turn by the characteristic concept “foils (swords)”.

On the left side, we see the morphological aspect, starting from “sword-like”, and penetrating through the concepts “swords” and “foils (swords)”. “Fencing swords” takes a hybrid position. Further, we show a constitutive aspect, the “wooden swords”. In the hierarchy here, wooden swords appear as weapons, and to our knowledge at least in medieval Japan, wooden swords were in use as respectable weapons. Alternatively, one could put wooden swords as a formal/constitutive hybrid under “sword-like objects”, and distinguish between “wooden swords (toys)” and “wooden swords (weapons)”. From this picture, we can easily recognise how the generalisation or specialisation has a direction towards the formal, the constitutive, and the functional aspect. A similar graph can be seen in Pustejovsky (1995:145).

“Weapons” is an established, “historical” term and concept, which is purely functional. Obviously, we cannot assign any prototypical shape to “weapon”, as we do to “sword”. As such, we regard it as descriptive and not characteristic. However, in general, nonformal terms are more frequently compounds than the others. In particular concepts which complement gaps in the lattice of traditional terms in order to cover relevant feature combinations compounds with terms from the participating aspects can be noticed, as expected from the faceted classification schemes presented above. The difference of our position here is that we cannot replace a part of the hierarchy by purely descriptive compounds as propagated by current indexing systems, because characteristic terms and traditional terms penetrate such hierarchies. A solution to embed these in a comprehensive way must be found, in a scheme of dynamic feature combination, as indicated in this paper.

6. Polydeykis, a practical experience

POLYDEUKIS is a general purpose, ontological Thesaurus aimed at consistently classifying cultural and archaeological terms. It is being elaborated by the Directorate of the Archive of Monuments of the Greek Ministry of Culture in co-operation with ICS-FORTH. It has reached its first state of completion, where the basic principles can be verified and further refinement can be determined. Its appellation was coined to honour Polydeukis, a second-century BC Alexandrian who authored a Lexicon of terms of the Attic dialect. The specific activities of the Directorate of Monuments have instigated a thematologically multi-faceted logical organisation of cultural and Archaeological terms contained in Polydeukis. Indeed, the Directorate handles an archive of extensive and multifaceted data concerning cultural monuments. These are the material traces of coherent human activity spanning over several millenniums from a geographical area considerably larger than the Modern Greek State. The influences exerted by the Hellenic culture over other cultural environments with which it held converse and still does, are very broad and remarkably fertile.

Consequently, all objects under investigation are characterised by an exceptional variegation of terminological designation and distinction into characteristic concepts. The analytical overview of concepts in Polydeukis follows an anthropocentric logical structure, that is to say, it observes closely, albeit up to a certain degree, the relationship which man (as a living being) has with his fellow humans, and with the creations created by himself or nature. Every human being, for instance, is a biological entity designated by his parents (creators), the time and place of birth, name, physiognomy and bodily form, personality, biological structure, as well as his activities, and the ambient wherein he lives and works.

In analogy, the logical organisation of Polydeukis establishes a conceptual anthropocentric structure for every type of object/ monument, to which all characteristic historic terms designating and specifying it are linked through a system of guide terms. The analysis of characteristic concepts conducted so far yielded the formulation of a relatively repetitive conceptual structure for their every set.

Let us illustrate the above by the example of the concept “temple” (in Greek “Ναος”). This concept has two designated uses according to Polydeukis: a descriptive and a characteristic historical one. The descriptive concept “temple” designates every single building of religious function, regardless of the religion or the religious tradition it belongs to. As such, it is a specialisation of a morphological descriptive term “building”, and the functional term “object from religious life”. “Temple” is further specialised into characteristic historic terms by the particular religious tradition it belongs to, i.e., “sanctuary, temenos, temple, dominicium, mosque” (temple appears again in its narrower sense). Associated with it are characteristic concepts of temple “elements”, functional parts like: “altar, nave, opisthodomos, diakonikon, prothesis, choir, narthex, royal doors, mihrab, minbar” and structural parts like: “capital, architrave, chancel barrier, frieze, pediment, mukarnas, etc.”. In a similar way, each characteristic concept is embedded in a conceptual structure capturing its social, functional, morphological and stylistic significance with the corresponding terminology.

The conceptual structure underlying all terms contained in Polydeukis also follows the above-mentioned anthropocentric model. The distinctions of concepts in Polydeukis are purely methodological: they are meant to convey not the levels of the very nature of things, but the levels of man’s cognition according to the specific standpoint by which he searches and understands the

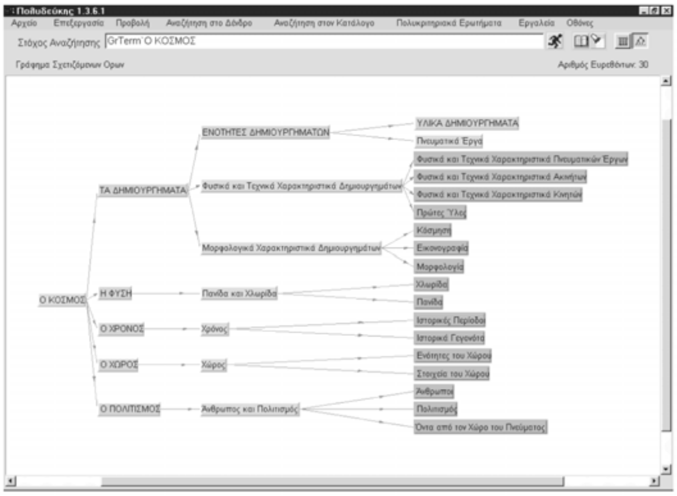

Figure 3: Screen dump from Polydeykis, "The Kosmos".

nature of things. For that purpose, Polydeukis uses the following major facets (see figure 3):

- Kosmos, the world as subject
- Living Nature, as historical subject
- Culture and Civilisation
- Space
- Time
- Creations, the man-made world:

  Material creations
  Conceptual works
  Associated concepts: Stylistic, physical and technical characteristics.

Hereby "Kosmos" includes other facets as subjects of representations. The "Living Nature" is organised by the relation to man, habitat and feeding habits, rather than genealogy. The man-made objects (as a formal aspect "ειδος"), their associated concepts (constitutive aspect) and relevant social/functional terms from the Culture and Civilisation facet are combined systematically between each other in two levels of hierarchy of descriptive terms into a kind of regular grid. The combinations are announced group wise by guide terms like "functional parts of immobile objects". Under these, the characteristic terms are introduced, and the hierarchy continues on demand with the multiple criteria. E.g. "immobile objects" are split into "simple constructions", "buildings", "built complexes", "housing areas", "open spaces", "other immobile objects", whereas activities are split into "religious life", "burial rites and functions", "private life", etc. (see figure 4). The combination of "buildings" and "religious life" leads to, for example, "temple", as seen above.

This kind of hierarchical layout allows for the formation of a common structural environment for the descriptive and the historical/characteristic level. According to such an organising principle of the Polydeukis material, the higher we are in the hierarchy of terms, the more we function on a level of descriptive categorisation. On the other hand, the more we descend from the terms at the top, the more we are apt to encounter characteristic historic terms. The main characteristic of this superintending view of facets is that the underlying synthetic structuring of terms follows a flexible policy of conceptual integration which is capable of registering the whole spectrum of variegations of concepts whether descriptive or historic/characteristic. Furthermore, it allows for the formation of broader or differentiated descriptive categorisations for the same characteristic concepts.

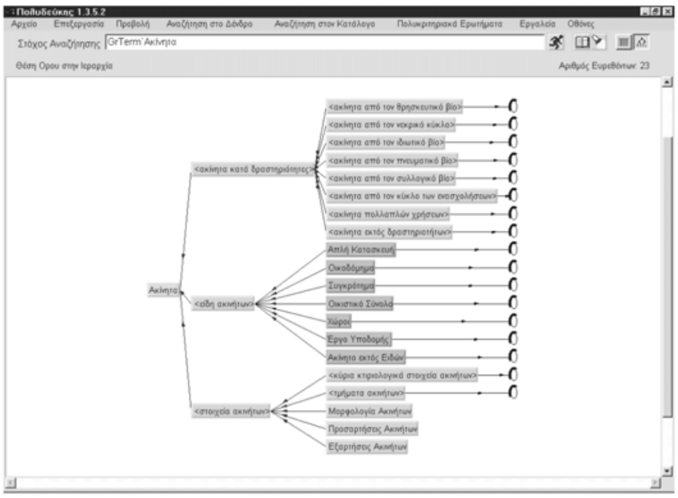

Figure 4: Screen dump from Polydeykis, immobile objects hierarchy.

The first experience with users is quite encouraging. The systematic structure of the upper levels and the explicit relation to descriptive concepts provide a remarkable orientation for the user, and we got enough positive responses. Characteristic terms seem to "find a natural place" in the hierarchy. On the other hand, the creation of combinations leads to a kind of "combinatorial explosion", and is not easy to be maintained. Also, the introduction of guide terms at all levels makes the presentation "heavy". Speaking with Ranganathan, we have successfully ignored the limitations on the notational plane, overcome the verbal plane by introducing compound terms, and satisfied the idea plane. Now, the next step is to introduce a suitable notation, and to improve the compound term creation.

7. Conclusions

The experience with Polydeykis, the analysis of AAT guide terms and the linguistic theory of qualia structures makes us confident, that the analysis of a thesaurus by placing object terms into a limited number of aspects and the respective logical and notational organisation of thesauri is feasible and can considerably improve user orientation, precision of classification and reasoning during retrieval. We have introduced the notion of the "descriptive" versus "characteristic" concept, which seems to be a viable distinction able to account for certain phenomena in the classification. An initial distinction into "modern" and "historical" was soon abandoned, as we could find all kinds of transitions between historical and modern terms with respect to their significance. We would be interested to learn if similar notions have been developed in philosophy or cognitive science.

We have encountered limitations of the current notational practice. We believe, that these can be overcome by mechanisms of virtual pre co-ordination based on Description Logic that we have conducted our first experiments with (Ntoas 1999). Eventually also new modes of graphical presentations need to be developed. However, several open questions about thesaurus structuring remain, for example:

1. Should there be a relation between certain qualia and the level of abstraction, or between descriptive and characteristic terms in a practical thesaurus? E.g. do functional terms tend to be more abstract than characteristic or formal ones (as psychological studies suggest)?
2. Are the qualia of common sense language the same to those of a specialist vocabulary? Are the latter similar, a part of the latter or can they be completely alien, e.g. in some technical fields?
3. A formalisation of multi-typed hyponymy for thesauri, i.e. broader/narrower term relations specific to qualia, should be developed.
4. To what detail should sense distinction in thesauri be made in order to account for historic concept shifts? Where is the optimal precision/complexity?

Finally, this approach may be useful to merge or make the transition easier between multiple thesauri in networked environments, as it can give a better account for reasons and aspects of broader/narrower term distinctions.

References

BECHHOFER, S., GOBLE, C.A, 1999. *Classification Based Navigation and Retrieval for Picture Archives*. IFIP WG2.6 Conference on Data Semantics, DS8 , Rotorua, New Zealand.

BORGIDA, A., 1995. Description Logics in Data Management. *IEEE Transactions on Knowledge and Data Engineering*, 7(5): 671-682.

CONSTANTOPOULOS, P. and DOERR, M., 1995. Component Classification in the Software Information Base. In Nierstrasz, O. and Tsichritzis, D. (eds.), *Object-Oriented Software Composition*. Prentice-Hall.

CHEN, H., LYNCH, K., BASU, K. and NG, D., 1993. *Generating, Integrating, and Activating Thesauri for Concept-based Document retrieval*. IEEE Expert, 8(2): 25-34.

[DMOZ] DMOZ Open Directory Project http://dmoz.org/

[DL] More information can be found for instance on http://www.ida.liu.se/labs/iislab/people/patla/DL/index.html

DOERR, M. and FUNDULAKI, I., 1998. *A proposal on extended interthesaurus links semantics*. Heraklion - Crete, Greece: FORTH, Institute of Computer Science - Technical Report FORTH-ICS/TR-215.

DOERR, M. and CROFTS, N., 1999. *Electronic Esperanto: The Role of the Object Oriented CIDOC Reference Model*. Proc. of the ICHIM'99, Washington, DC.

[Encarta] Microsoft Corporation: "*Mircrosoft Encarta Reference Suite 2000*", CD SET X04-99037, 1993-1999.

GETTY ART HISTORY INFORMATION PROGRAM AHIP, 1994. *Introduction to the Art & Architecture Thesaurus*. Published on behalf of The Getty Art History Information Program, Oxford University Press, New York.

HAZEWINKEL, M., 1996. *Enriched Thesauri and their Uses in Information Retrieval and Storage*. First Delos Workshop, An Overview on Projects and Research Activities in Digital Library Related Fields. INRIA - Sophia Antipolis. *ERCIM Workshop Proceedings* - No. 97-W001.

[MDA97] Museum Documentation Association, English Heritage & Royal Commission on the Historical Monuments of England, "*mda Archaeological Objects Thesaurus*", 1997, http://www.open.gov.uk/mdocassn/archobj/archcon.htm

MICHALSKI, R.S., 1993. *Beyond Prototypes and Frames: The Two-tiered Concept Representation*. In Van Mechelen, I., Hampton, J., Michalski, R. and Theuns, P. (eds.), *Categories and Concepts, Theoretical Views and Inductive Data Analysis*, Academic Press, San Diego.

MILLER, A., GEORGE, BECKWITH, RICHARD, FELLBAUM, CHRISTIANE, GROSS, DEREK and MILLER, K., 1993. *Introduction to WordNet: An On-Line Lexical Database*.

MURRAY-RUST, P. and WEST, L., 1998. Knowledge, Language and Semantics: XML and VHG (TM). *ASLIB Managing Information* vol.5, no.4: 34-36.

NTOAS, D., 1999. *Economy and consistency in Thesauri*. Heraklion - Crete, Greece: FORTH, Institute of Computer Science -Technical Report FORTH-ICS-TR-262.

PINKER, S., 1994. *The Language Instinct*. New York: W. Morrow and Co.

POLLITT, S., 1997a. *Interactive Information Retrieval based on Faceted Classification using Views-Knowledge Organisation for Information Retrieval*. Proceedings of the 6th International Study Conference on Classification, University College, London 16-19 June 1997. http://www.hud.ac.uk/schools/cedar/dorking.htm

POLLITT, S., 1997b. *The key role of classification and indexing in view-based searching*. IFLA '97 Copenhagen Aug 31 - Sept 3 1997. 63rd IFLA General Conference Booklet 4. Section on Classification and Indexing Session 95 Paper 009-CLASS-1-E.

PRIETO-DIAZ, R. and FREEMAN, P., 1987. *Classifying Software for Reusability*, IEEE Software: 6-16.

[PrivCom] Private Communication with Tanya Szraiber, British Museum.

PUSTEJOVSKY, J., 1995. *The Generative Lexicon*. MIT Press. ISBN 0-262-16158-3.

RANGANATHAN, S.R., 1965. *A descriptive account of Colon Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science.

RECTOR, A., BECHHOFER S., GOBLE, C., HORROCKS I., NOWLAN, W. and SOLOMON, W., 1997. *The GRAIL Concept Modelling Language for Medical Terminology*. THE GALEN Project, Medical Informatics Group, Department of Computer Science, University of Manchester, England.

[SHIC93] The SHIC Working Party, "Social, Historical and Industrial Classification", The Museum Documentation Association, Cambridge 1993, ISBN 0 905963 91 1.

SVENONIUS, E., 1989. *Design of Controlled Vocabularies*. Encyclopaedia of Library and Information Science, New York, Marcel Dekker: 82-109.

[TEI] Text Encoding Initiative (http://www.uic.edu/orgs/tei), Chapter 13, “Terminological Databases” http://www.uic.edu/orgs/tei/p3/doc/p3te.txt.

VICKERY, B.C., 1960. Thesaurus - A new word in documentation. *Journal of Documentation,* 16:4: 181-89.

WELTY, C., and JENKINS, J., 1999. Welty, Formal Ontology for Subject. *Knowledge and Data Engineering.* 31(2): 155-182. Elsevier.

WELTY, C., and IDE, N., 1999. Using the Right Tools: Enhancing Retrieval from Marked-up Documents. *Journal of Computers and the Humanities*, 33(1-2). Kluwer.

SOERGEL, D., 1995. The art and architecture thesaurus (AAT): A critical appraisal. *Visual Resources*, X: 369-400.