Journal of Digital Information, Vol 1, No 8 (2001)

Semantic Problems of Thesaurus Mapping

Martin Doerr
Institute of Computer Science, Foundation for Research and Technology - Hellas,
Heraklion, Crete, Greece
Email: martin@ics.forth.gr

Abstract

With networked information access to heterogeneous data sources, the problem of terminology provision and interoperability of controlled vocabulary schemes such as thesauri becomes increasingly urgent. Solutions are needed to improve the performance of full-text retrieval systems and to guide the design of controlled terminology schemes for use in structured data, including metadata. Thesauri are created in different languages, with different scope and points of view and at different levels of abstraction and detail, to accommodate access to a specific group of collections. In any wider search accessing distributed collections, the user would like to start with familiar terminology and let the system find out the correspondences to other terminologies in order to retrieve equivalent results from all addressed collections. This paper investigates possible semantic differences that may hinder the unambiguous mapping and transition from one thesaurus to another. It focusses on the differences of meaning of terms and their relations as intended by their creators for indexing and querying a specific collection, in contrast to methods investigating the statistical relevance of terms for objects in a collection. It develops a notion of optimal mapping, paying particular attention to the intellectual quality of mappings between terms from different vocabularies and to problems of polysemy. Proposals are made to limit the vagueness introduced by the transition from one vocabulary to another. The paper shows ways in which thesaurus creators can improve their methodology to meet the challenges of networked access of distributed collections created under varying conditions. For system implementers, the discussion will lead to a better understanding of the complexity of the problem.

1 Introduction

Terminological resources are increasingly important for information retrieval in wide area networks, for retrieving documents by querying databases and metadata employing controlled vocabularies. In particular, thesauri which organize terms and associated concepts in the form of simple semantic networks become important tools for searching through the rapidly growing electronic information flood. There is growing interest in developing automated intermediaries to negotiate the differences between controlled vocabulary schemes so that a user can use a familiar set of terms to search collections using other vocabulary schemes.

The paper discusses the effect of thesaurus mapping on the vagueness in retrieval from a theoretical and logical point of view, separate from the effects of the relation of the thesaurus to the collection it addresses. Therefore, it makes ideal assumptions for the latter without going into any detail. It assumes that a certain collection is tightly connected to a terminological resource, which is given in the form of a thesaurus containing typed relations between terms and concepts, as defined by Foskett (1997). By "tight connection" we mean, in one case, that data values or classification terms are restricted to this thesaurus and consistently used. In that case, a query with some term should retrieve exactly the objects meant (this may actually require the expansion of a query term into its narrower terms). Let's refer to this case as the *controlled vocabulary situation*. Alternatively, the collection may use free text or free keywords. In this case, what we mean by "tight connection" is that a search-aid thesaurus exists that approximates the (expert) language used in the collection and the associated concepts. In this case, we assume that this thesaurus provides better results for that resource than any other. Let's refer to this as the *free text situation*. Finally, we are not concerned with documents only, but with museum object descriptions and others as well. The term *objects* is used here to cover all collection objects, not just text documents.

1.1 Related Work

There is a vast literature and established practice about creating thesauri for the purpose of information description and retrieval (for example, publications of the Getty Research Institute, the British Arts and Humanities Data Service (AHDS), and various national standards). Thesauri are designed as an agreement and compromise on a set of shared terms for common concepts. Even though the agreement on term definitions is sought over large communities, thesauri even in the same domain differ significantly (Krause 2000), and there is limited interoperability between tools and digital resources employing different thesauri.

To resolve the incompatibility between different terminology resources, research has initially concentrated on attempts to unify thesauri by merging, e.g. Mili and Rada (1988), Mannino *et.al.* (1988), Rada and Martin (1987). The Unified Medical Language System (UMLS) merges concepts from some 50 sources into a metathesaurus, which retains links to its original sources. It is probably the largest merging effort undertaken so far (Nelson 1999). Two problems have turned out to be the most difficult:

- First, differences in term semantics, semantics of hierarchical relations and term overlap can render the simple combination of concepts from two sources impossible. Resolutions to this problem are usually semiautomatic, see e.g. Constantopoulos and Sintichakis (1997), and can become fairly expensive.
- Second, controlled vocabularies are often associated with a large installed base of systems using them, such that migration to a new set of terminology and relations may be virtually impossible; for example, with subject headings used by national libraries (Chan 2000, Landry 2000).

Recently the focus has been more on creating systems that provide a transition from one thesaurus or subject heading scheme to another. Two paradigms exist:

- thesaurus correlation (ISO5964 (1985), Getty Information Institute (1996), Doerr and Fundulaki (1998a)), where substituting concepts are sought
- thesaurus federation (Kramer *et al.* 1997), where lead-in terms to related concepts are sought.

Nikolai *et al.* (1999) propagated both approaches together. The German CARMEN project has begun to correlate different German thesauri used to index social science literature, using intellectual and statistical methods simultaneously. In the MACS project (Landry 2000), the Consortium of European National Libraries (CENL) is currently investigating the feasibility of correlating the subject headings of all European languages. When terminology in different natural languages is combined, the result is often referred to as "multilingual thesaurus." In the Term-IT project it was found that creating multilingual thesauri and combining different thesauri in the same language belongs to the same class of problem (supported by Krause 2000). More about "multilingual thesaurus" correlation is included in section 2.2.

The problem of combining subject headings and classification systems, such as Library of Congress Subject Headings (LCSH) and Dewey Decimal Classification (DDC), has been investigated by Chan (2000) and Vizine-Goetz (1998). Chan writes: "How to combine the salient features of a rich vocabulary like LCSH and the structured hierarchy found in classification schemes such as LCC and DDC to improve retrieval of networked resources remains a fertile field for research and exploration", and advocates the harmonization of vocabularies. Chan refers to the efforts of the Library of Congress to make the LCSH more useful for networked access by improving its faceted structure, the principles of term construction, and rigorous term relationships. In this paper, we argue for development in the same direction.

1.2 Intellectual versus Automatic Term Correlation

Term correlations may be intellectually created, as by MACS, the French Ministry of Culture (see Merimee), the HEREIN Project, and CARMEN (Krause 2000). Krause calls intellectually created term correlation lists "cross-concordances." Alternatively, they may be created by statistical methods or even by neural networks, as in CARMEN and Vizine-Goetz (1998). Chen *et al*. (1996) use a concept space approach to create thesauri automatically and to traverse between two thesauri from different biological subdomains. They report a considerable increase of recall in a cross-domain search experiment, but differences between the links provided by the algorithm and those given by experts. Statistical and neural network methods do not easily allow interpretation of the intellectual nature of a given link, if at all. They are far cheaper, however, and can detect relations of which humans are unaware. The ultimate precision is usually low. As Chan (2000) puts it: "the tension between quality and quantity has never been keener." As this paper deals with semantic problems, we do not consider statistical methods, even though we are convinced that the future lies in the coordinated combination of intellectual and statistical methods, as the CARMEN project and others do.

An interesting point in Chen *et al*. (1996) is the report that the term associations users made in cross-domain searching were context-driven: "...Based on our protocol analysis, we found that several contexts for these similarities and differences existed, including, two genes were identified by similar (or different) experimental strategies; their cellular structures had similar (or different) composition; ... genes manifested similar or different phenotypes; genes or proteins had similar or dissimilar sequences (homology) or contained similar motifs or domains; proteins or genes performed similar (or dissimilar) functions;... ." This paper contributes to a better understanding of such phenomena. Even though semantic heterogeneity of terminological resources has frequently been referred as a problem, a systematic analysis of its intellectual basis and structure has not been carried out. Krause (2000) writes: "... the information market over the past twenty years... mainly views the development in distributed databases, user interfaces and the Internet as technological improvements and problems of standardization, without addressing the conceptual challenges involved."

1.3 Thesaurus Mapping in One Domain

We regard thesaurus mapping as the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent. It is a central process for merging thesauri, metathesaurus and cross-concordance construction, and thesaurus switching. This paper investigates the problems of finding appropriate equivalents, in particular focussing on issues related to polyhierarchies and the relationships between compound and non-compound terms.

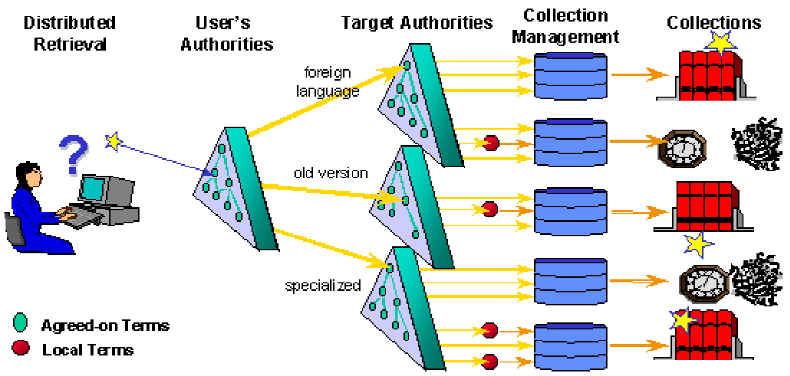

Figure 1. Scenario using correlated thesauri

The following assumes a general scenario (Figure 1) of a user addressing different digital collections using a particular thesaurus of choice, which is mapped to thesauri in other languages, to more specialized vocabularies, or to different versions of the thesaurus. We adopt the notion of a two-step process from Krause (2000a), which separates the vagueness introduced by thesaurus mapping (step one) from that introduced by the relationship between the user query and the document (step two). To separate these effects intellectually, each thesaurus is assumed to be consistent with the indexing of one or more collections, i.e. a correct user query to one collection through its own thesaurus yields full recall and high precision. Such ideal conditions can exist, for instance, in databases indexing museum objects with a thesaurus about the physical object types. We further assume that the set of objects in all collections is basically of the same nature and from one domain, to exclude another source of vagueness.

Under these conditions, we attribute the remaining heterogeneity between different thesauri to (see Doerr 1996, p.3):

1. **Different word use,** due to different natural languages, the chosen language level, not semantically justified decisions in the selection of descriptors, and the degree of post- or pre-coordination of terminology. For example, the following terms all describe the same species: chaffinch (English), Buchfink (German), fringilla coelebs (scientific). Such differences are readily apparent when comparing terminology of thesauri in the same language with overlapping domains, such as the Art & Architecture Thesaurus from the Getty Foundation and the NMR Monument Type Thesaurus from English Heritage (formerly RCHME, in the following "NMR"), both in English.
2. **Different coverage,** due to different states of development, different scope and varying user needs. In particular, thesauri often develop some topics in far more detail than others. For example, something found under "dolls, Hopi" in

one thesaurus may be found under "Kachina" in another. Some place name thesauri may cover places only down to the country level, while another includes lower administrative areas and communities.
3. **Different semantics,** due to different conceptualizations. This occurs typically between thesauri from different languages, but it may also be due to different aspects of classification. For example, does "architecture" and its narrower terms denote types of buildings or the designing of buildings? Is "museum" a building or an organization?
4. **Different semantic relations,** often due to the enforcement of monohierarchies, but also due to different classification aspects. When polyhierarchies are not permitted, placement of terms that have two or more likely hierarchies (e.g. "chemical geology","decorative weapons") is based on considerations that are often not documented; two monohierarchical thesauri are likely to make different decisions about placement, which may not primarily be based on concept semantics.

We shall give several examples from the Art & Architecture Thesaurus (AAT) Getty AHIP 1994, because we have been able to study it in detail and because it is an impressive example of a large, context-free thesaurus created with strict editorial principles ("context-free" means without a commitment to a specific application). Years after it began, it is now also a source to study what could be done better, a fact that does not diminish its extraordinary value (see also Soergel 1995). Other examples are taken from the intellectually created mapping of the Merimee "Thesaurus Architecture" to the AAT and the English Heritage Thesaurus (NMR) in the 1997 edition, as used in the AQUARELLE project and now being carried on in the HEREIN Project.

Section 2 presents well-known ideas of concept-based thesauri in order to clarify several notions. First, we define the kind of mapping we mean to clarify the differences to other approaches. Second, we discuss distinct classes of "multilingual thesaurus", which is used in a fuzzy way in literature. Third, we refine the current notions of equivalence expressions, as given in ISO5964, in order to conform to certain logical requirements. Based on that background, we make a novel proposal for a methodology of mapping that allows for controlling vagueness in cross-thesaurus retrieval. This becomes increasingly relevant, as several projects are beginning to create such mappings on a large scale. Section 3, studies effects that may either impede the definition of equivalences between terms or between hierarchies, or impede the exploitation of semantic relations in the target thesaurus for query expansion. Chan (2000) regards thesauri as "a query expansion device", a virtue that should be preserved through mapping.

2 Application of Concept-based Mapping

Even though the ISO *Guidelines for the establishment and development of multilingual thesauri* (ISO5964) were published in 1985, it is only now that such mappings are being attempted on a larger scale. Since its publication, no specific methodology has been proposed about which terms should be correlated with equivalence relations and which not in the lack of exact equivalence, and how this would affect the query or information retrieval quality of the pair of correlated thesauri as a whole.

2.1 Concept-based mapping

In this section and some of the following we talk about sets in the mathematical sense. By *objects* we do not mean only documents. It may be anything in an electronically registered collection: a potsherd, a stool, a palace, an image, and a text. By *sets* we usually refer here to sets of such objects, typically defined by the sharing of one or more common properties. We cannot go into more details about Description Logic and similar theories here, and we can mention only the basic idea:

Under certain assumptions, preferred terms, so-called "descriptors", can be identified with concepts. Each concept in turn can be identified with the **intention** of a **set** of objects. In the sequence, we can transform the mapping problems into a mathematical problem about sets, i.e. terms are identified with the sets of objects they correctly classify (see Doerr and Fundulaki 1998 for details). "Correctly" is a question of user convention, and we assume that users can in general positively decide which term is correctly applied and which not. This assumption provides an absolute measure to compare concepts in thesauri even between multiple languages. As long as objects in a large enough database are classified in a well-defined way with two thesauri in parallel, set-relations between the concepts of both thesauri can be approximated automatically (as by Amba *et al*. 1996). Any inconsistencies can then be reduced to human errors. Such assumptions are well known and basic to Description Logic (DL), e.g. Baader *et al*. (1992), Borgida (1995), DL Web site, and implicit in many thesauri describing physical objects. In practice, not all subset relations may have been expressed in a thesaurus and term interpretation can be context dependent in a complex way, as will be discussed later. To make a clear distinction from statistical or neural network methods, let's define "concept-based mapping". The principles are:

1. A term is mapped to the associated set of objects which it correctly classifies (like the "interpretation function" in Description Logic).
2. The associated set of a broader (narrower) term (BT/NT) of some term is a superset (subset) of the associated set of the latter. In terms of DL, the broader term "subsumes" the narrower.
3. Some kinds of related term (RT) relationships can be identified with roles in the sense of DL, in particular the part-whole relation (BTP) and functional relations.
4. The **mapping** between two terms is defined through the **set-relations** of their associated sets.

In this definition we use the relation notations BT, NT, RT, and BTP of ISO 2788. This definition of mapping is stricter than the term "cross-concordances" used by Krause (2000). It must be stressed that concepts are interpreted by descriptor use and not by comprehension of the term itself. Reuse of a thesaurus in a different context may change interpretation and require a redesign of the hierarchical relations (BT/NT), a complication often overlooked (see section 3).

2.2 About Multilingual Thesauri

Often any kind of relations between terms from thesauri in different natural languages are referred to as translations. In our opinion, translation in the proper sense differs from the concept-based mapping and cross-concordances in significant ways. In the AQUARELLE project, Dachelet (1997) proposed distinctions between different kinds of multilingual thesauri. To clarify the differencese we define the following classes of "multilingual thesaurus":

1. **Translated thesauri.** A thesaurus, where each concept is optimally interpreted in words of another or multiple languages, to allow speakers of those languages to understand better and use the concepts of this thesaurus more effectively. Note that such translations are in general not established indexing terms of the target language.
2. **Correlated thesauri.** An aggregate of multiple thesauri consisting of established indexing terms (concepts) of the respective user communities, and a set of concept-based mappings between the concepts from the different thesauri of that aggregate. The mappings serve as **replacements** of the original terms in queries sent out against multiple databases. Each database uses one of the thesauri of such an aggregate (see e.g. the AQUARELLE project). The

replacement is done in order to obtain **equivalent results** from all databases to the degree possible. Note that the mapped concepts are in general not good translations of each other, because they do not interpret the other concept, but the associated sets of the mapped concepts contain common objects. This definition can be relaxed to weaker kinds of correlations. It is consistent with ISO 5964, but obviously not restricted to the case of different natural languages.

3. **Interlingua.** A thesaurus made out of concepts that are created by fusing each cluster of similar concepts from different social groups into a new concept (by analogy with the solution in machine translation, see e.g. Hutchins (1995)). In this process, one term from each user group is attached to the new concept as the identifier to be used by this group. The interlingua provides the sharing of concepts between social groups, e.g. as a legal basis used by the European Commission like the EBTI. Note that the interlingua may not contain any of the original concepts of any user group; it contains a set of compromises to remove interpretational differences. Its concepts may again be translated and correlated to other thesauri.

Figure 2 illustrates symbolically the three types of multilingual thesauri. It alludes to the example of the correlations created by the French Ministry of Culture between the Merimee "Thesaurus Architecture" and the English Heritage thesaurus (NMR). For example, Merimee's *tennis* = NMR's *tennis court* **AND** *tennis club*; *megalithe* has as narrower equivalence *standing stone*. We have added hypothetical translations and a hypothetical interlingua. Sometimes a system of well-defined concepts from one group is adopted as a whole by a user group of another language. For example, the Library of Congress Subject Headings are used in our Greek university library and in many libraries of other smaller language groups. Similarly, there exist translations of the *American Art & Architecture Thesaurus* to Dutch and Spanish, and others are planned. In such cases the concepts in the translated thesaurus can become the interlingua, with the benefit that it contains at least one's original concepts. In this paper, we are interested primarily in correlated thesauri, which are also the aim of ISO 5964.

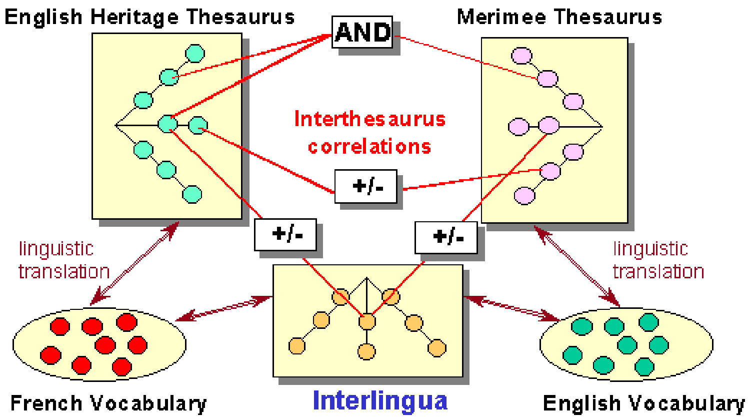

Figure 2. Demonstrating different notions of a multilingual thesaurus in one context

2.3 Equivalence Expressions

Equivalence expressions similar to those in ISO 5964 are used with increasing frequency for thesaurus mapping: in the Merimee "Thesaurus Architecture", CARMEN, MACS, the HEREIN Project and others. Based on the idea of concept-based mapping and on the argument that correlated thesauri should serve equivalent retrieval results across systems employing different terminological resources, we have proposed that the expressive power of the mapping should be at least equivalent to the expressive power of the search paradigm. Otherwise, the user could express better queries in each target system than the mapping mechanism could provide. Doerr and Fundulaki (1998) investigated the mapping equivalent to the Boolean expressions (AND, OR, NOT) foreseen by the Z39.50 protocol. We found that a slight extension to ISO5964 is sufficient to achieve equivalence expressions with the expressive power of Boolean queries. For that purpose, we interpret the equivalence expressions of ISO5964 as concept-based mappings, i.e. as set relations of the associated sets of objects. This seems to be justified by the Venn-diagram-like illustrations in the ISO5964. We make the following interpretations and extensions:

- "**partial equivalence**" should become "**broader equivalence**" (is subset of) or "**narrower equivalence**" (is superset of)
- **"exact equivalence**" is interpreted as "same set as"
- "**inexact equivalence**" is interpreted as "overlaps with"
- "**single to multiple equivalence**" should become "*** equivalence"** to "compound" where "compound" is a Boolean expression of target terms with AND, OR, NOT and "*****" is either "**exact**" or "**broader**" or "**narrower**".

The Getty Information Institute (1996) proposed the symbolism of a broader ("<") and narrower (">") equivalence, AND combinations ("+"), and OR combinations ("&"), as does the on-going HEREIN project. Other Boolean expressions have not yet been proposed. Boolean expressions are interpreted as intersections, unions and complements of the associated sets. The UMLS Metathesaurus, due to its tighter coupling, uses implicitly exact, broader and narrower equivalences, as well as explicit term combinations ("Associated Expressions") using AND combinations.

So far, these equivalence expressions provide a means to express initial query terms in terms of any target thesaurus. Obviously, any Boolean combination of terms in the initial query can be converted into a Boolean combination of target terms (see below). Figures 3 and 4 demonstrate the semantics of equivalence expressions with Boolean compounds. An example for Figure 3, the French term *bergerie* has the exact equivalence to "*sheep barns* **OR** *sheep folds*" in the AAT. The first common broader term of both terms in the AAT is *single built works*. The obvious broader term *animal housing* is the broader term in AAT only to *sheep folds*, probably because of its monohierarchy design (see section 3.3). For Figure 4, please see the list in the Appendix. In Figure 4, the dotted circle on the right-hand side indicates where the approximated concept would appear in the target hierarchy, under the assumption that the BT relation expresses subsumption.

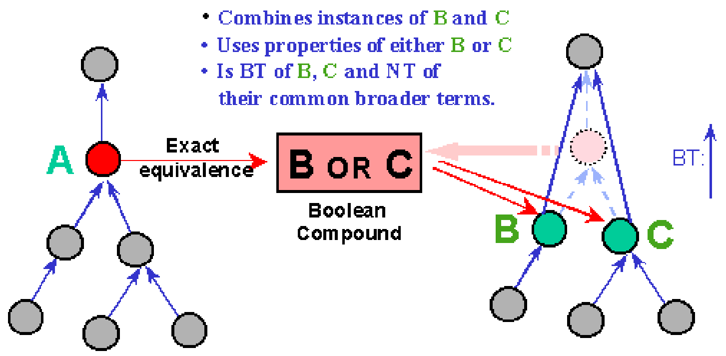

Figure 3. Demonstrating equivalence to **OR** combinations of terms

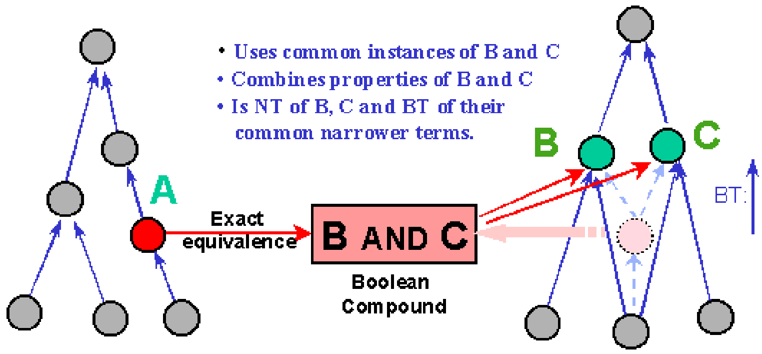

Figure 4. Demonstrating equivalence to **AND** combinations of terms

2.4 Methodological Aspects

Above, the arguments are constrained to the *controlled vocabulary* situation: concept-based mappings, derived from ISO5964, used to create correlated thesauri. Under these restrictions the effect of the following methodological arguments can be evaluated theoretically. These restrictions are realistic. For example, the HEREIN project will connect databases about material cultural heritage with thesauri correlated in the style of ISO5964. The precision of manual object classification is precise in the way assumed in section 2.1. Other obvious cases of precise classification are the use of place name authorities (gazetteers) like the *Thesaurus of Geographic Names* (TGN) and cultural period authorities. Professional users in cultural heritage administration (as investigated in the AQUARELLE and Term-IT project) and many other disciplines require stricter standards for recall and precision than general users seeking information on the Internet. The optimization of the recall/precision ratio usual in information retrieval does not satisfy the needs of a statistical survey. Chan (2000) also refers to the requirement for recall and precision as distinct: "Subject access tools are used to enable optimal recall ... to enable optimal precision...". In the *free-text* situation, basically the same arguments presented below should hold, but the effect will not be so explicit because several factors introduce additional vagueness. The question of when the effect of more elaborate correlations vanishes in the vagueness coming from other sources is interesting, but beyond the scope of this paper. The same holds for the question of whether the effort to create mappings intellectually or semi-automatically is affordable or not. We are satisfied here with the fact that people are increasingly undertaking that effort.

To illustrate the relevance of the following, Table 1 presents some statistics about the carefully produced equivalence expressions from the 1997 editions of the French Merimee "Thesaurus Architecture" to the AAT and NMR thesaurus. Table 2 compares the frequency of equivalence expressions among these three thesauri.

Table 1. Characterizing three thesauri

Thesaurus	Domain	Language	Hierarchies	Pre-coordination
Merimee	Western architecture	French	mono	high
ATT	Art and Western architecture	English	mono	low
NMR	Western architecture	English	poly	high

Table 2. Distribution of equivalence relation types in the Merimee mappings

Total number of Merimee descriptors:	**1336**	**100%**
Number of terms with equivalence to AAT:	**795**	**59%** of all terms
Number of exact equivalence to AAT:	**687**	**85%** of all equivalences to AAT
Number of partial equivalence to AAT:	**119**	**15%** of all equivalences to AAT
Number of **OR** combinations to AAT:	**26**	**3%** of all equivalences to AAT
Number of **AND** combinations to AAT:	**196**	**25%** of all equivalences to AAT
Number of terms with equivalence to NMR:	**735**	**55%** of all terms
Number of terms with equiv. to AAT and NMR:	**634**	**48%** of all terms
Number of exact equivalence to NMR:	**596**	**78%** of all equivalences to NMR
Number of partial equivalence to NMR:	**165**	**22%** of all equivalences to NMR
Number of **OR** combinations to NMR:	**86**	**11%** of all equivalences to NMR
Number of **AND** combinations to NMR:	**8**	**1%** of all equivalences to NMR

In the mapping to the AAT, AND combinations mainly reflect post-coordination rules. Of the 199 so-called AND combinations used by Merimee to map to the AAT, at least 174 turned out to be role restrictions rather than true AND combinations (see section 3.3). They are listed in the Appendix. NMR is far more detailed and pre-coordinated, therefore OR combinations dominate, with **up to 6!** terms combined. The AND combinations to NMR follow the logic of Figure 4. As one team has created the above equivalences, the differences must be attributed to the nature of the target vocabularies and not to differences in the practice of the editors.

Some 40% of the Merimee terms are not mapped, with no indication at all which terms they relate to in the other thesauri. For example, the French term *EDIFICE FUNERAIRE* has no equivalence to the NMR. Its narrower terms *MAUSOLEE* and *OSSUAIRE* have one equivalence each: *mausoleum*, *ossuary*. Obviously, the broader term is *mausoleum* in NMR: *funerary site* could be a broader equivalence of *EDIFICE FUNERAIRE*, but that has not been declared. Probably the editors felt that it would be "too far". This is an example of the proposal illustrated in Figure 5. Table 2 illustrates the complexity of thesaurus mapping, and shows that mappings not created with a well-founded methodology for networked information retrieval do not provide the necessary qualities, as will be discussed in the rest of this paper. Moreover, we think that providing such quality would not require a much greater effort than that invested in a mapping like that presented above.

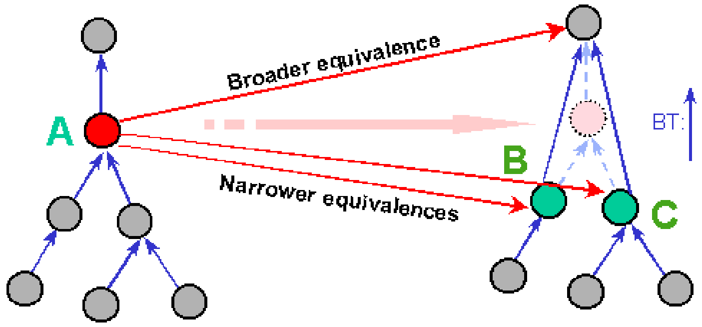

Figure 5. Demonstrating semantics of term inclusion

Now consider the question of whether an optimal mapping is possible. If someone uses an arbitrary set of equivalence expressions for correlation, the existence of an equivalence link for some query terms is not guaranteed. So a query using such a term would simply fail against the respective target. Even if an equivalence expression exists, the relation between the intended and the actual query with replaced terms is *a priori* unpredictable. Equivalence relation creation in a larger environment (Hutchins 1995) causes some "combinatorial explosion", therefore "switching languages" are proposed as intermediaries. The results of such subsequent mappings are even more undefined. We therefore propose to approximate concepts of a source thesaurus systematically by confining them within the nearest broader and narrower concepts of the target thesaurus (Figure 5). The idea can be stated in the following rules:

1. If a concept c of thesaurus A has no exact equivalence to any concept of thesaurus B then: for this concept c at least one broader equivalence and at least one narrower equivalence to some appropriate concepts in thesaurus B should be declared (see Figure 5).
2. The above **broader** equivalence should be **minimal**, i.e. there should be no other term combination in B which is broader than concept c and narrower than the given broader equivalence.
3. The above **narrower** equivalence should be **maximal**, i.e. there should be no other term combination in B which is narrower than concept c and broader than the given narrower equivalence.
4. The user has the chance to specify if the transformed query is going to preserve recall or precision.

If Rule 1 holds, we consider this a "complete mapping." If Rules 1, 2, and 3 hold, the correlation is considered to be an "optimal mapping." Obviously, choosing a broader equivalence for a single term in a Z39.50 protocol access profile element will preserve recall, i.e. return a set containing the relevant items plus some non-relevant ones. Negation turns things over: the narrower equivalence will preserve recall, and the broader precision. In practice, both recall preservation and precision preservation are needed; e.g. recall for statistics (under the assumption, that uncorrelated items can be sorted out in a second step) and precision for rapid discovery of relevant items.

Rule 1 defines a notion of "completeness" of the mapping, that is, if Rule1 holds, the replacement of any query term is possible. Of course, it may be impossible to find any narrower equivalence, in which case we use the empty concept ("bottom" in DL). In this case, the query returns no result, which is consistent. In the case of negation, however, it would return the universal concept ("top" in DL), that is, the whole target base. This is a case to be prohibited. Even though, such a situation can be tolerable in a query refinement cycle, as the actual intermediate results may not be transferred. In rare cases it may even be impossible to find a broader concept, in which case one must map to the universal concept. This situation could be avoided if thesaurus providers agree to share some high-level concepts. In addition, other AND combinations as they appear in a typical user query may "absorb" the universal concepts and return reasonable results. If the result of a translated query would be the whole target base, the user should be informed and given the choice to cancel the query. We see here an area for **pragmatic solutions** in the framework of application development. Note also that concept inclusions propagate without problem through multiple intermediate translations.

The above qualitative reasoning holds in this form only for simple conjunctive queries. In general, the problem can be reduced to a query containment problem, if the database schema plus terminology are interpreted as a schema altogether. For the latter, elaborate theories about the complexity and decidability of query containment exist (Calvanese *et al.* 1998 on PODS), which have to be applied on a case-by-case basis. Calvanese *et al.* (1998) on KR'98 present a fairly general, unified framework for information integration from heterogeneous sources.

Rules 2 and 3 above define a notion of an "optimal mapping" in the sense that no closer mapping can be found with these kinds of expressions. If one actually chooses Boolean combinations rather than only the primitive concepts for the term inclusion, things may become algorithmically complicated and can go beyond the capacity of normal expert insight. Not even typical Description Logic implementations as CLASSIC or FaCT answer such questions. On the other hand, if the mapping is close to optimal, the vagueness control still exists. Here is an area for further applied research.

Finally, the use of Boolean compounds poses some more methodological questions. Whereas other equivalence expressions can be read (anti-) symmetrically (reciprocally),e.g. *narrower equivalence* reads reciprocally as *broader equivalence,* a Boolean compound can **not** be easily interpreted in the opposite direction (see Figures 3 and 4). Thesaurus or DL tools don't even normally indicate which concepts are used in a Boolean compound. Our laboratory has implemented a research prototype (Ntoas 1999) on top of the thesaurus management system SIS-TMS (Doerr and Fundulaki 1998a), which indexes expressions containing Boolean operators and role restriction ("restrict [p,C]") and relates them to their immediate broader and narrower terms. Future research could address the question of the degree to which Boolean compounds or more complex DL expressions of some mapping can be exploited to calculate equivalence expressions in the opposite direction.

Summarizing, under the given assumption the proposed mapping methodology allows for the propagation of any query to collections classified with thesauri that have been mapped to one another. The query will **not fail** for any given query term, except if the whole database should be returned. As concepts do not precisely match, the results **cannot** always be equivalent. Rather, following the choice of the user, a (**smallest possible**) **larger** result or a (**largest possible) smaller** result can be returned, which could be further refined through post-processing steps. Such results can be used for statistical purposes.

For the free text situation (see section 1), *inexact equivalence* as defined in ISO5964 should be quite useful in balancing recall and precision. I guess, however, that the appropriate generalization of an *inexact equivalence* for full-text retrieval is a correlation based on associated relevance weights. We shall not follow this subject further here.

All of the above is based on the ideal assumption that the correlated thesauri follow the same rules and that their BT relations are complete, consistent, and follow the same logic. If this is not the case, some consistency checks can be carried out. At least it can be verified if manually or automatically derived equivalence relations won't cause cycles with the given BT relations on either side, i.e. if some concept seems to include one of its broader concepts. In some cases the search for equivalences may reveal missing additional BT relations on either side, as *sheep barns* BT *animal housing* above. The rest of this paper is devoted to thoughts about the reasons for inconsistencies between hierarchies, the problems appearing in reality.

3 Heterogeneities of the Hierarchical Structure

If two correlated thesauri use subsumption hierarchies (or IsA relationship) and declare explicitly all direct subsumption relations between their primitive (non-compound) concepts, many nice applications can be done. The transitivity of subsumption allows expansion of query terms into their narrower terms to arbitrary depth, in particular by correlated concepts of another thesaurus and their narrower terms. This allows *switching* use from one thesaurus to another, e.g. to a more specialized one. Thus a general *high-level* thesaurus can be federated with a series of application-specific thesauri. Further, the subsumption relations between **all** terms of two thesauri can be calculated from a complete mapping in the above sense, and eventual logical inconsistencies can be reduced to human errors and eliminated. In practice, however, term hierarchies often (1) do not express subsumption, (2) are ambiguous, or (3) do not express all immediate subsumption relations. Some reasons and possible solutions are analysed below.

3.1 Hierarchical relation without subsumption

Traditionally, thesauri were printed books, and the hierarchies were used as an association mechanism to lead users most effectively to a concept for which he or she does not know the term. The sequencing into book pages does not foster the use of polyhierarchies. Hence, thesaurus hierarchies were more like **decision trees** than semantic relations. With computers, representation restrictions become obsolete (Welty and Jenkins 1999). Fascinating in this context are Ranganathan's (1965) classical considerations about the obstacles the "notational plane" causes to the development of the "ideal plane". The traditions from editing printed books are not easily overcome, however. So often any hierarchical relation is messed up with subsumption, as they are equally useful for user guidance. In our opinion, user guidance and semantic relations are not all the same and should **coexist** in the same thesaurus.

ISO 2788 still regards the **part-whole** relation (**BTP**) as a kind of Broader Term relation, whereas e.g. the AAT Editorial Manual (1998) already regards them as a kind of Related Term (RT, "Code 2B"), and no longer as subsumption. Another example is the **inclusion of geographical areas** in place name thesauri (e.g. the *Thesaurus of Geographical Names* (TGN)). Obviously, France isA Europe does not hold. Different hierarchical relations hold for **temporal intervals** and cultural **periods,** even though there are still few examples of thesauri about periods. The CIDOC CRM ontology (Doerr and Crofts 1999) refers to these four relations as *forms part of*. These four relations can be explained by their extensions to different related sets, i.e. sets of points on the surface of earth, in an object, on the time-line, and in space-time. Hence they inherit the partial order relation from the subsumption of the respective sets, form (poly)hierarchies, and are therefore frequently mistaken for *broader terms*. Transitivity does not extend among them, however (see Motschnik (1993) about limited transitivity between different part-whole relations), and therefore they cannot be mingled. Nevertheless, expansion of query terms, e.g. from object types into their parts, can be useful if explicitly required.

Another source of confusion are the semantic relations within a set of derived concepts, *parallel hierarchies* as described by Soergel (1995) or the DL role restriction (Baader *et al*. 1992). For example, even though "*Greece* IsA *Europe*" does not hold, "*Greek person* IsA *European person*" does hold. In this case, *Greek person* is interpreted as *Person.who lives in: Greece*. *Who lives in* can be seen as a DL role, here restricted to *Greece*. Another example: even though "*bridge construction* IsA *bridge*" does not hold, "*book about bridge construction* IsA *book about bridges*" may be regarded as valid. The use of terms in a specific database field may hide a concept derivation, e.g. when object names are used as subjects or place names as nationalities. Editors may introduce in their thesaurus the subsumption relations correct for that use, i.e. of the derived concepts. Out of context those can be completely wrong.

In particular, **subject headings** often refer to **physicalobjects** (e.g. objects in museum collections), but their hierarchies cannot be directly used for object classification, causing frequent disputes between librarians and museum curators. Characteristically, the Getty Information Institute has rearranged large amounts of terms from the LCSH (Library of Congress Subject Headings) into the AAT, a thesaurus mainly about physical objects (its *Object* facet, except for the hierarchy *information objects*). We regard it as worthwhile to investigate under which conditions associations such as the above can cause subsumption relations in their derived concepts. Related to this is Welty and Jenkins' (1999) thorough study on modeling subjects. We would not expect the opposite, i.e. that concept derivation from a subsumption hierarchy may not preserve this hierarchy in the derived concepts. At least theoretically it should not happen (Ntoas 1999). Subject terms used for library cataloging are sometimes interpreted as applying to books which cover the breadth of the term. For example, *biology of mammals* would be used to denote books about the biology of **all** mammals, rather than **some** mammals. In this interpretation, narrower terms are **not** subsumed.

Finally, thesauri such as the "dmoz" project, which lets users act freely, end up with associations motivated by any contextual link, and may even contain cycles and other relationships that break thesaurus rules. For example, both "Top: Arts: Classical Studies: Journals" and "Top: Arts: Classical Studies: Academic Departments" are declared as narrower terms of "Top: Arts: Classical Studies" (DMOZ).

Summarizing, there are hierarchical relationships used in thesauri which are not subsumptions. They have to be clearly marked as being of different nature so that correct reasoning can be done. Otherwise, *screws* may be taken for *cars*, *villages* for *nations*, *Andorra* for a *continent*, etc. The use of terms in a specific database field may hide a concept derivation, and out of context hierarchies made for that use could be wrong. Therefore, the assumed semantics of hierarchical relations should be made clear before thesauri are correlated and should be communicated as thesaurus **metadata or clear notations for the relationships**.

3.2 Context-induced ambiguity

Terms and concepts often reveal a polysemy which is disambiguated by the context in which they are used. English, for instance, is full of so-called *homonyms* or *contrastive ambiguity* (Pustejovsky 1995), like: *orange (color)* and *orange (fruit)*; *pink (color)* and *pink (vessel)*; *column (architectural element)* and *column (text arrangement)*. Even though some older thesaurus maintainers insist on word-based hierarchies, concept-based hierarchy organization prevails now in computer science. Consequently, the concepts have to be disambiguated. For example, in the AAT, a domain determinator like *color* in *orange (color)* disambiguates the concept. The actual word (e.g. *orange*) can be attached as a *non-preferred term* or *synonym* to all possible meanings. Word Net (Miller *et al*. 1993), for example, uses many-to-many relations between words and concepts, the most consistent approach to represent the real relation.

So far, the problems of homonymy seem to be solved by this so-called *sense enumeration* (Pustejovsky 1995). Each sense of a term represents a concept independent of contextual influence, and subsumption hierarchies can be designed independent of use. Homonymy is a language-specific feature, i.e. the different senses of one term in one language are normally translated into different terms in another language.

There are, however, the more subtle cases, which Pustejovsky calls *complementary polysemy*, an expression of the dynamic power of the concept formation behind our languages. For example, is *door* an object or an opening? Is *neck* a part or a place on a body? Is *school* an organization or a building? These terms are typically translated one-to-one into other languages for the same set of meanings. Pustejovsky introduces the notion of **qualia**, the different aspects that may cause a word to change meaning in context. I have the impression, that this polysemy may be intrinsic to the concept itself. He talks about the *Qualia Structure* of nominals, which he analyses in the following main 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, color, position)
- Telic: purpose and function of the object (purpose an 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, artifact, natural kind, causal chain)

This analysis has a striking similarity with criteria for the term specialization we found from an empirical study (Doerr and Kalomoirakis 2000) in the guide terms of the AAT. Terms about man-made objects, for example, systematically seem to have **functional (telic)**, **morphological (formal)**, and **constitutive** aspects. We have the feeling that a **few** *qualia* dominate in each application. If this is the case, they can be formalized and their consequences identified and communicated as thesaurus **metadata**.

A clear distinction between concept and term, as in the case of contrastive ambiguity, cannot be made, and Pustejovsky regards sense enumeration as impractical. Three problems arise from that:

1. Often thesauri are not made with the full breadth of application in mind. Consequently, one or other aspect may be neglected. For example, the AAT regularly defines classifications by form and by function, but it seldom relates more specific concepts to both categories. Figure 6 demonstrates the necessity of multiple generalizations using the example of *foils*.
2. A thesaurus made to capture a specific aspect (e.g. in the AAT, the morphological aspect dominates) may provide insufficient or even wrong hierarchies if used under another aspect. Figure 6 symbolizes with different colors how each aspect may give rise to different kinds of narrower term relations. This subject may deserve further investigation.
3. The coverage of a concept may vary under these aspects. For example, the functional versus morphological aspects of a *sword*: is a children's wooden toy sword a *sword*? There were quite functional wooden swords in Japan. Shops are full of samurai sword **imitations**. In the Archeological Museum of Heraklion, there is a presumed part of a **non-functional** Minoan sword, pretty sharp, which seems to have been an instrument for dangerous artistic exercises. And finally, there are **sword-like** letter openers.

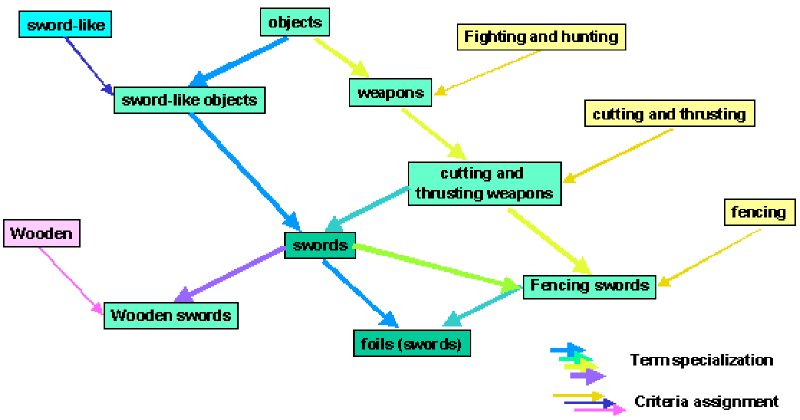

Figure 6. Multiple broader terms under multiple aspects

To the best of our knowledge, the problems arising from complementary polysemy for thesaurus design has not yet been studied. Under the above considerations, it is useful to make the *qualia* of the use of a thesaurus explicit and to add such characteristics to thesaurus **metadata**. For example, a *hammer* in the morphological sense, as archeologists would classify items, has nothing in common with a *steam hammer*. This may even provide incorrect broader terms for the functional aspect an engineer needs, who sees both concepts closely related as types of *impact devices*.

On the other side, thesaurus editors should systematically take into account other aspects of use and identify the **additional** broader term relations the other aspects require, as long as they are not contradictory. Maybe a better solution would be to make BT relations aspect-specific, as suggested by the colors in Figure 6. A similar situation is shown by Pustejovsky (1995) on p. 145. Of course, we are aware that this may become quite labor intensive if no automatic methods are found. The whole topic leaves many questions open. Chen *et.al*. (1996) on the context-driven associations cited in section 1 confirms our impression that these problems are highly relevant for thesaurus mapping. Such difficulties seem to be often ignored by technicians or taken as unavoidable vagueness of human argument.

3.3 Missing subsumption relations

Besides reasons of different contexts of use, we see the enforcement of **monohierarchies** and post-coordination rules as the major reasons for missing subsumption relations. As mentioned above, monohierarchies were preferred as long as the predominant medium for thesauri were books. A nice example is the *colorant* hierarchy of the AAT. Look at the position of *crimson lake* and *carmine* in this hierarchy in Figure 7.

```
- <materials by function>
- -colorant
- - -pigment
- - - - <pigment by color>
- - - - - red pigment
- - - - - - organic red pigment
- - - - - - - crimson lake

- <materials by function>
- - colorant
- - - <colorant for dye and pigment>
- - - - carmine
```

Figure 7. Position in hierarchy of the AAT terms "crimson lake" and "carmine"

There are three inherent aspects: functional form (pigment, dye, lake...), appearance (red, blue, brownÃƒÂ¢Ã¢'Â¬Ã,Â¦), production or provenance (artificial, organicÃƒÂ¢Ã¢'Â¬Ã,Â¦). The above sequencing structure seems to have placed the two terms arbitrarily. One could as well start with color or provenance. *Carmine* does not appear at all under its characteristic color, because priority was given to other prominent features (*carmine* is used for dye **and** pigment). The scope note of *crimson lake* states: "Deep, transparent, ruby-red lake pigment with bluish undertone, made from kermes, a natural dyestuff of insect origin; *carmine*, a better pigment introduced in the 16th century, became its chief competitor. MAYER." Hence *carmine* is a red organic pigment, but does not appear under that term and is far away in hierarchy from a very similar one.

Imagine a thesaurus federation: a thesaurus may have a leaf node of organic colorant and we would like to employ the AAT as the source for more specific terminology. Even though all concepts are in principle in the AAT, and there is **no** difference in conceptualization, we cannot continue from organic colorant to narrower terms in the AAT. In a student project, we have created an experimental colorant hierarchy, where each colorant was put directly under three broader terms: functional, color, and provenance terms. The complete result is difficult to show graphically, but browsing is effective, because on descending one branch or the other one comes down to the correct end and no possible broader term is missing.

This brings us to the other point: **post-coordination**. Obviously it is quite inefficient to define all combinations of terms like *artificial inorganic red pigment*, *synthetic organic green pigment*, etc. Therefore, thesauri like the AAT and the German subject headings *Schlagwortnormdatei* (SWD) use rules to combine terms dynamically. For reasons of simplicity, only a "+" and "&" signs are used, both heavily overloaded with different interpretations. For example, "*factories + grinding*" in the AAT means a "factory which does grinding"; i.e. nothing more than a *mill*. The term *mill* has been sacrificed to post-coordination (it is a *decoordinated subject* in the AAT terminology), as are many other useful terms, as can be seen in the Appendix from the Merimee mappings. See also Soergel's (1996) extensive analysis of such problems in the AAT. In current practice post-coordination suffers from three problems:

1. Unclear semantics and no consistency control.
2. Under the current convention, post-coordinated terms are leaf-concepts, because no mechanism is foreseen to relate another established term to a post-coordinated compound, as a narrower term or other.
3. No indication of which parts of hierarchies it makes sense to combine.

Hence it is fairly difficult to reconstruct the missing terms and their broader terms, a major obstacle to thesaurus mapping and federation - but neither can precoordination be seen as a solution, because it overloads thesauri. Problems 1 and 2 are appropriately solved by DL. Languages like GRAIL (see also Rector *et.al.* 1997) provide a user-friendly syntax. Problem 3, however, is an open issue. Rich DL implementations do not, for obvious reasons, allow browsing through all possible concept combinations.

On the assumption that it makes sense for specific parts of hierarchies to be post-coordinated with specific relative roles, Ntoas (1999) designed a mechanism in which the user can declare that a specific concept or subhierarchy can be refined by restriction of a specific role to another subhierarchy. For example, *factories which do* can be declared to be valid for a subhierarchy of processes, like *factories which do grinding,* etc. In the sequence, the user can browse through the *virtual hierarchy* induced by the subsumption properties of the respective processes. Further, explicitly declared natural concepts like mill will appear at their natural position in the virtual hierarchy, and narrower terms of mill can be added, which is impossible in the AAT. Boolean combinations, which do not define or lead to a natural concept, were not included in the browsing. They play a minor role in natural concept formation and are easily handled and understood by users.

Roles were taken from concepts in the CIDOC CRM (Doerr and Crofts 1999) ontology and from roles we found to be implicit in AAT terms. The semantic relations of the UMLS Semantic Network are another source of relevant roles, not only for the medical domain. It seems that a few relatively generic roles may actually be sufficient for most cases. The mechanism was verified with term combination from the equivalence expressions between the 1997 editions of the French Merimee "Thesaurus Architecture" and the AAT listed in the Appendix. For example, the compound "*factory* & *owner's* & *houses*" is interpreted as "*houses* which has owner: *Person*, which is owner of: *factory*"; *wood* & *roofs* as "*roofs* which consists of: *wood*"; *umbrellas* & *factories* as "*factories* which produce: *umbrellas*", etc.

The above examples and Figure 6 demonstrate that post-coordination is useful at all levels of hierarchy, but that there must be a mechanism to embed natural concepts in post-coordinated schemes. We regard mechanisms simulating hierarchies of post-coordinated concepts as necessary to mediate effectively between pre- and post-coordination in thesaurus mapping and federation.

3.4 Summary

Section 2 showed that a suitable methodology to create thesaurus mappings can provide well-defined global recall and precision qualities for transitions between thesauri, that have so far not explicitly been considered. Thus we made assumptions that are realistic but often not present. In this section, we have studied effects that may undermine those assumptions. Some can be avoided, either by better awareness of the thesaurus providers or by specific reasoning services. In other cases, information about implicit assumptions can help avoid comparing incomparable structures. Therefore, we have repeatedly proposed that certain thesaurus characteristics be documented in metadata, to avoid semantic heterogeneity conflicts and to facilitate interoperability of reasoning mechanisms. These characteristics are:

1. The use and notation of different kinds of hierarchical relations, which are not transitive among each other.
2. The use of subsumption relations, not for the concept itself but for use-dependent derivatives.
3. The aspect (qualia) under which subsumption hierarchies are created. This point needs further investigation.
4. The use of monohierarchies, provision of complete polyhierarchies, incomplete polyhierarchies
5. Details of post-coordination rules, which hierarchies can be combined, and the roles in use.

The Networked Knowledge Organization Systems (NKOS) group is an informal user organization devoted to the discussion of the functional and data model for enabling the interoperability of knowledge organization systems, such as classification systems, thesauri, gazetteers, and ontologies. For that purpose, is has defined a metadata format for KOS, the NKOS Registry. It provides a virtually complete description of the technical and administrational characteristics of a KOS, except that data about cardinality constraints are not required for all types of relationships. Aspects of use as above are not analysed in detail. We propose to extend this metadata format by a suitable formulation of the five criteria above. The detail of semantic analysis we postulate here may appear to be too labor-intensive to pay off. This is, however, not a concern of a qualitative study like this. If one regards, for example, the incredible progress traditional dictionary writing has made by semiautomatic methods, I am quite confident that the near future will provide reasonable solutions. Already a scientific community has begun to concentrate on the issue of Ontology Learning (see conclusions of the OntoWeb Workshop). Therefore, we regard a good understanding of the intellectual problems of thesaurus integration as quite beneficial.

4 Conclusions

Many things can be done to bring forward information integration with thesauri. There is still a large gap between practitioners and scholars on one side, and theoreticians in knowledge representation and system engineers on the other. Whereas the practitioners administer the domain knowledge, the others have the technology to improve its handling. It is not easy for the practitioners to understand and appreciate the potential and limitations of technology, and often the theoreticians do not show particular understanding for intellectual problems in practice that do not directly conform with their models. We see a need for increased interdisciplinary empirical studies and **verification** of theoretical results, both to demonstrate the **utility** of theory to practitioners and to identify their **limits** and need for better theoretical understanding.

The results of several excellent implementations of thesaurus federations seem to have remained relatively unevaluated in terms of the real quality of concept mediation achieved. Some technology providers seem to see their task end at the point of installation and optimize their systems to work with any thesaurus. As we have tried to make clear in sections 2 and 3, we believe that **thesaurus creators** (scholars, experts and practitioners) have a responsibility to **improve** their **methodology** to meet the challenges of advanced technology (e.g. the completeness of hierarchies) and **technologists** have a responsibility to understand the complexity of the problem (e.g. contrastive polysemy). If this were to happen in a coordinated way, we could soon achieve a new quality of applications.

We believe that some concrete steps could be undertaken:

1. We see theory and practice that has advanced enough to implement query transformation services that preserve recall and precision in a controlled manner. To that end, a certain methodology of thesaurus creation and protocols for connecting to knowledge resources on the Internet should be agreed on. The interface between translation *services* incorporating thesaurus mappings with clearly defined logical properties should be defined.
2. Methodological specifications for thesaurus providers should be developed that conform to respective technical requirements. Among these are: isA-semantics of broader terms, completeness of broader terms, synonym creation for concept identification in free text, explicit aspects for hierarchy creation (functional, morphological, etc.)
3. Metadata for thesauri and the services employing mappings should sufficiently define the properties and methodological principles needed for coherent, dynamic information integration services.

We also see scope for applied research in the connection with formal ontologies (in the sense of knowledge representation models as semantic networks, KL-ONE-like data models or Description Logic) with thesauri. These are:

1. The identification of basic roles for concept formation (like the process-product-producer relation, *using*, *made for*, etc.). We believe this needs the cooperation of ontologists, linguists and empirical studies in terminological resources. These would improve handling of post-coordination, in terms of use, interpretation and consistency control. Important lessons can be learned from the formal handling of medical terminology (see e.g. Galen Project and GRAIL and from the UMLS Semantic Network).
2. The handling of transitions between terms and values in multiple data fields in information integration; e.g. a record with "type = *wooden sword*" versus a record with "type=*sword* and material=*wood*."
3. The provision of a high-level framework for integrating multiple knowledge resources under fundamental categories like *Actors*, *Physical Objects*, and *Events* (see e.g. the CIDOC CRM) beyond the medical domain.

Finally, we see a need for more research in the understanding of dynamic concept formation and the conditions under which conceptual hierarchies can be compared, cooperatively used, or merged.

Acknowledgements

I wish to express my thanks to the reviewers of this paper, and particularly to Linda Hill, for their valuable comments, which greatly helped to improve its quality.

References

Amba, S., N. Narasimhamurthi, K.C. O'Kane and P.M. Turner (1996) "Automatic linking of thesauri". In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Konstanz:Hartung-Gorre, pp. 181-187

AQUARELLE Telematics Applications Programme, Information Engineering Sector, Project IE 2005, "Final Report", http://aqua.inria.fr

Baader, F., H-J. Burckert, J. Heinsohn, B. Hollunder, J. Muller, B. Nebel, W. Nutt and H. Profitlich (1992) *Terminological knowledge representation: a proposal for a terminological logic*, DFKI Report, DFKI, Saarbruecken

Borgida, A. (1995) "Description logics in data management". *IEEE Trans. on Knowledge and Data Engineering*, 7(5):671--682

Calvanese, D., G. De Giacomo and M. Lenzerini (1998) "On the decidability of query containment under constraints", In *Proc. of the 17th ACM SIGACT SIGMOD SIGART Sym. on Principles of Database Systems* (PODS'98), pp. 149-158

Calvanese, D., G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati (1998) "Description logic framework for information integration", In *Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning* (KR'98), pp. 2-13

Chan, Lois Mai (2000) *Exploiting LCSH, LCC, and DDC To Retrieve Networked Resources Issues and Challenges*, Library of Congress, December 19
http://lcweb.loc.gov/catdir/bibcontrol/chan_paper.html

Chen, Hsinchun, J. Martinez, T. D. Ng, and B. R. Schatz (1996) "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System". *Journal of the American Society for Information Science*, Vol. 47, No. 8, August
Copy available at http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html

Constantopoulos, P., M. Sintichakis (1997) "A Method for Monolingual Thesauri Merging". *Proc. 20th International Conference on Research and Development in Information Retrieval, ACM SIGIR*, July, Philadelphia, PA, USA

Dachelet, R. (1997) "Multilingual querying and multilingual thesauri in Aquarelle", Technical Report, INRIA-Aquarelle, March

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

DMOZ Open Directory Project, http://dmoz.org/

Doerr, M. (1996) "Authority Services in Global Information Spaces:A requirements analysis and feasibility study", Technical Report FORTH-ICS/TR-163, February

Doerr, M., I. Fundulaki (1998) "A proposal on extended interthesaurus links semantics". Technical Report FORTH-ICS/TR-215, FORTH, Institute of Computer Science, Heraklion - Crete, Greece

Doerr, M. (1998) "Effective Terminology Support for Distributed Digital Collections". In *Sixth DELOS Workshop, Preservation of Digital Information*, Tomar, Portugal, June

Doerr, M., I. Fundulaki (1998a) "SIS - TMS: A Thesaurus Management System for Distributed Digital Collections", *Proc. 2nd European Conference, ECDL'98*, September, Heraklion, Crete, Greece

Doerr, M., N. Crofts (1999) "Electronic Esperanto: The Role of the Object Oriented CIDOC Reference Model". *Proc. ICHIM'99*, Washington, DC, September

Doerr, M., D. Kalomoirakis (2000) "A Metastructure for Thesauri in Archeology". *Proc. CAA2000*, Lubljana, April

EBTI: A short description of the EBTI (European Binding Tariff Information) Thesaurus can be found in: http://www.bjl.be/2_3_1.htm

English Heritage, National Monuments Record (2000) *NMR Monument Type Thesaurus,* June 19
http://www.rchme.gov.uk/thesaurus/mon_types/default.htm

Foskett, D.J. (1997) "Thesaurus", In *Readings in Information Retrieval,* edited by K. Sparck Jones and P. Willet (Morgan Kaufmann), pp. 111-134

Getty AHIP (1994) *Introduction to the Art & Architecture Thesaurus*. Published on behalf of The Getty Art History Information Program (New York: Oxford University Press)

Getty Information Institute (1996) *Guidelines for Forming Language Equivalents: A Model Based on the Art&Architecture Thesaurus*, International Terminology Working Group (for copies contact Murtha Baca, mbaca@getty.edu).

The **HEREIN** Project http://www.european-heritage.net/fr/Thesaurus/Contenu.html

Hutchins, W. J. (1995) "Machine Translation: A Brief History". In *Concise history of the language sciences: from the Sumerians to the cognitivists,* edited by E.F.K.Koerner and R.E.Asher (Oxford: Pergamon Press), pp. 431-445

ICOM/CIDOC Documentation Standards Group (1998) : "*CIDOC Conceptual Reference Model*", http://www.ville-ge.ch/musinfo/cidoc/oomodel/index.htm

ISO 2788-1986 (1986) *Documentation - Guidelines for the establishment and development of monolingual thesauri,* International Organization for Standardization, Ref. No ISO 2788-1986

ISO 5964-1985: (1985) *Documentation - Guidelines for the establishment and development of multilingual thesauri*, International Organization for Standardization, Ref. No. ISO5964-1985

Kramer, R., R. Nikolai, and C. Habeck (1997) "Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies". In *International Journal on Digital Libraries* (1), 122-131

Krause, J. (2000) "Virtual libraries, library content analysis, metadata and the remaining heterogeneity". *Proc. ICADL 2000, the 3rd International Conference of Asian Digital Library,* Seoul, Korea

Krause, J. (2000a) "Information Systems for Social Science Research. A perspective from Information Science". In *Proceedings of the Symposium Information System for Social Sciences*, Mannheim, Germany

Landry, P. (2000) "The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD)". *Classification and Indexing Workshop, 66th IFLA Council and General Conference*, Meeting No. 181 http://www.ifla.org/IV/ifla66/papers/165-181e.pdf

Mannino, M.V., S. B. Navathe, and W. Effelsberg (1988) "A Rule-Based Approach for Merging Generalization Hierarchies". *Information Systems*, 13(3):257-272

MERIMEE, "THESAURUS ARCHITECTURE" for the indexing of complexes, buildings and built works described in the national database "Merimee" about the French Heritage http://www.culture.gouv.fr/documentation/thesarch/pres.htm

Mili, H., R. Rada (1988) "Merging Thesauri: Principles and Evaluation". *IEEE Transactions On Pattern Analysis and Machine Intelligence*,10(2):204-220

Miller, A. G., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller (1993) *Introduction to WordNet: An On-Line Lexical Database*

Motschnik-Pitrik, R. (1993) "The Semantics of Parts Versus Aggregates in Data/Knowledge Modelling". In *Proc. CAISE'93*, Paris, June (Berlin: Springer-Verlag), pp. 352-361

Nelson, S. J. (1999) "The Role of the Unified Medical Language System (UMLS) in Vocabulary Control". *CENDI Conference on Controlled Vocabulary and the Internet*
http://www.dtic.mil/cendi/presentations/nelson.ppt (Powerpoint file)

Nikolai, R., R. Kramer, M. Steinhaus, B. Felluga, and P. Plini (1999) "GenThes: A General Thesaurus Browser for Web-based Catalogue Systems". In *Proceedings of the Third IEEE Meta-Data Conference*, Bethesda, Maryland, April

NKOS, Networked Knowledge Organization Systems/Services
http://www.alexandria.ucsb.edu/~lhill/nkos

NKOS Registry (1998) Draft Set of Attributes, based on Contolled Vocabulary Registry developed by Linda L. Hill and Interconnect Technologies in 1996, with some modification, last revision: 7/30/98
http://alexandria.sdc.ucsb.edu/~lhill/nkos/Thesaurus_Registry.html

Ntoas, D. (1999) "Economy and consistency in Thesauri". Technical Report FORTH-ICS-TR-262, FORTH, Institute of Computer Science, Heraklion - Crete, Greece

OCLC, Online Computer Center (2000) *Dewey Decimal Classification*, Dublin, OH, USA (Forest Press)
http://www.oclc.org/fp

OntoWeb Workshop (2000) Semantic Web Project Proposal, organised by Dieter Fensel and Ying Dingat, Vrije Universiteit Amsterdam (the Netherlands), Dec. http://www.ontoweb.org/workshop/amsterdamdec8/index.html

Pustejovsky, J. (1995) *The Generative Lexicon* (MIT Press)

Rada, R., B. K. Martin (1987) "Augmenting Thesauri for Information Systems". *ACM Transactions on Office Information Systems*, 5(4)

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

Rector, A., S. Bechhofer, C. A. Goble, I. Horrocks, W. A. Nowlan, and W. D. Solomon (1997) "The GRAIL concept modelling language for medical terminology". *Artificial Intelligence in Medicine*, 9:139-171

Soergel, D. (1995) *The Art and Architecture Thesaurus (AAT): A critical appraisal*. Visual Resources, X, pp. 369-400

Term-IT Project home page http://www.mda.org.uk/term-it/

U.S. National Library of Medicine (2001) *2001 UMLS Metathesaurus,* January 12, section 2
http://www.nlm.nih.gov/research/umls/META2.HTML

U.S. National Library of Medicine (1998) *Fact Sheet UMLS Semantic Network,* February 19
http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

Vizine-Goetz, D. (May/June 1998) "Subject Headings for Everyone: Popular Library of Congress Subject Headings with Dewey Numbers". *OCLC Newsletter*, 233:29-33

Welty, C., J. Jenkins (1999) "Formal Ontology for Subject". *J. Knowledge and Data Engineering*, 31(2), September, 155-182

Z39.50, *ANSI/NISO Z39.50* or *ISO 23950*: *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification*
http://lcweb.loc.gov/z3950/agency/

Appendix: "&" combinations from the Merimee thesaurus to the AAT

Table 3. All "&" combinations from the Merimee thesaurus to the AAT in 1997

Count	French term	Relation	American term combination
1	ACADEMIE	exact equivalence	academy & buildings
2	AIRE DE CONCASSAGE	exact equivalence	crushing & floors
3	AIRE DE LAVAGE	exact equivalence	washing & floors
4	ARCHEVECHE	exact equivalence	bishop (prelate) & palaces
5	ARCHEVECHE	partial equivalence	archbishop & palaces
6	ARDOISIERE	exact equivalence	slate & quarries
7	BASSIN	exact equivalence	artificial & pools
8	BROSSERIE	exact equivalence	brush & factories

9	BUREAU	exact equivalence	factory & offices
10	CABLERIE	exact equivalence	electric cable & factories
11	CALVAIRE	exact equivalence	calvary crosse & monuments
12	CARTONNERIE	exact equivalence	cardboard & factories
13	CHAIRE A PRECHER EXTERIEURE	exact equivalence	exterior & pulpets
14	CHAMBRE DE COMMERCE	exact equivalence	board of trade & buildings
15	CHAMOISERIE	exact equivalence	chamois & factories
16	CHARPENTE EN BOIS	partial equivalence	wood & roofs
17	CHARPENTE METALLIQUE	partial equivalence	metal & roofs
18	CHARRETERIE	exact equivalence	cart & sheds
19	COLLATERAL	exact equivalence	side & aisles
20	CONCIERGERIE	exact equivalence	porter's & lodges
21	CONSERVATOIRE	exact equivalence	drama & schools
22	COURSIERE	exact equivalence	wall & passages
23	COUTELLERIE	exact equivalence	cutlery & factories
24	CRISTALLERIE	exact equivalence	crystal (leadglass) & factories
25	CROIX DE CIMETIERE	exact equivalence	cemetery & crosses
26	DEPENDANCE	exact equivalence	agriculture & outbuildings
27	ECOLE D'AGRICULTURE	exact equivalence	agricultural & schools
28	ECOLE D'ART	exact equivalence	art & schools
29,30	ECOLE DE DANSE	exact equivalence	ballet & schools **OR** dance & studios (work spaces)
31	<?dicule religieux chr?tien>	exact equivalence	christian &
32	<?difice du g?nie civil>	exact equivalence	civil engineering & buildings
33	EDIFICE RELIGIEUX CHRETIEN	exact equivalence	christian & religious buildings
34	ELEVATION INTERIEURE	exact equivalence	interior & elevations (building divisions)
35	ENCLOS FUNERAIRE	partial equivalence	churchyard & walls
36,37	ENSEMBLE CASTRAL	exact equivalence	castle & complexes **OR** chateau & complexes
38	ENSEMBLE D'INDUSTRIE ALIMENTAIRE	exact equivalence	food processing plant & complexes
39	ENSEMBLE D'INDUSTRIE CERAMIQUE	exact equivalence	ceramic & complexes
40	ENSEMBLE D'INDUSTRIE CHIMIQUE	exact equivalence	chemical & complexes
41	ENSEMBLE D'INDUSTRIE DU BOIS	exact equivalence	woodworking & complexes
42	ENSEMBLE D'INDUSTRIE DU PAPIER	exact equivalence	papermill & complexes
43	ENSEMBLE D'INDUSTRIE VERRIERE	exact equivalence	glass & complexes
44	ENSEMBLE DE CONSTRUCTION AERONAUTIQUE	exact equivalence	aircraft & complexes
45	ENSEMBLE DE CONSTRUCTION AUTOMOBILE	exact equivalence	motor vehicle & complexes
46	ENSEMBLE DE CONSTRUCTION MECANIQUE	exact equivalence	assembly plant & complexes
47,48	ENSEMBLE DE CONSTRUCTION NAVALE	exact equivalence	shipyard & complexes **OR** naval shipyard & complexes
49	ENSEMBLE DE FABRICATION DE MATERIAUX DE CONSTRUCTION	exact equivalence	building material & complexes
50	ENSEMBLE DE FABRICATION DES METAUX	partial equivalence	<metalworking plant>& complexes
51	ENSEMBLE DE PETITE METALLURGIE	partial equivalence	machine shop & complexes
52	ENSEMBLE DU GENIE CIVIL	exact equivalence	civil engineering & complexes
53	ENSEMBLE FORTIFIE	exact equivalence	fortification & complexes
54	ENSEMBLE FUNERAIRE	exact equivalence	funerary buildings & complexes
55	ENSEMBLE METALLURGIQUE	partial equivalence	<metalworking plant>& complexes
56	ENSEMBLE TEXTILE	exact equivalence	textile mill & complexes
57	ESCALIER INDEPENDANT	exact equivalence	freestanding & stairs
58	ETABLISSEMENT CONVENTUEL	partial equivalence	christian & religious communities
59	ETABLISSEMENT DE BAINS	exact equivalence	public baths & baths
60	ETABLISSEMENT NAUTIQUE	exact equivalence	boating & clubhouses
61	ETABLISSEMENT PORTUAIRE	exact equivalence	harbor & buildings
62	FAIENCERIE	exact equivalence	faience & factories
63	FECULERIE	exact equivalence	starch & factories
64	<fondations et sols>	exact equivalence	foundations (structural elements) & pavements (surface elements)
65	FOUR A CHANVRE	exact equivalence	hemp & ovens
66	FOURNIL	exact equivalence	bake oven & buildings
67	GANTERIE	exact equivalence	glove & factories
68	GARAGE	exact equivalence	automobile & repairshop
69	GAZOMETRE	exact equivalence	natural gas & storage tanks
70	GLACERIE	exact equivalence	mirror & glass & factories
71	HUILERIE	exact equivalence	vegetable oil & animal oil & factories
72	LAC DE JARDIN	exact equivalence	garden & lakes
73	LAVABO DE CLOITRE	exact equivalence	cloister & lavaboes
74	LOCAL SYNDICAL	partial equivalence	trade union & buildings
75	LOGEMENT DE CONTREMAITRE	exact equivalence	foremen's & houses
76	LOGEMENT PATRONAL	exact equivalence	factory & owner's & houses
77	LOGIS ABBATIAL	exact equivalence	abbots' & houses
78	MAISON AUX DIMES	exact equivalence	tithing & offices
79	MAISON MINIATURE	exact equivalence	miniature & houses

80	MILLIAIRE	exact equivalence	Roman & milestones
81	MONTJOIE	partial equivalence	pilgrimage & markers (monuments)
82	OBSERVATOIRE	exact equivalence	astronomical & observatories
83	OUVRAGE D'ART	partial equivalence	civil engineering & structures (single built works)
84	PARFUMERIE	exact equivalence	perfume & factories
85	PASSAGE COUVERT	exact equivalence	carriage & porches
86	PASSAGE D'ENTREE	exact equivalence	carriage & passages
87	PERCEPTION	partial equivalence	tax collectors' & offices
88	PLATRIERE	exact equivalence	plaster & factories
89	PUITS D'AERAGE	exact equivalence	ventilation & shafts (spaces)
90	RAFFINERIE DE PETROLE	exact equivalence	petroleum & refineries
91	RAFFINERIE DE SUCRE	exact equivalence	sugar & refineries
92	ROBINETTERIE	exact equivalence	plumbing hardware & factories
93	SALLE DU THEATRE	exact equivalence	theater & auditoriums
94	SAVONNERIE	exact equivalence	soap (organic material) & factories
95	SECHOIR A CHATAIGNES	exact equivalence	chestnut & drying sheds
96	SECHOIR A MAIS	exact equivalence	corn & drying sheds ;
97	SUCRERIE	exact equivalence	sugar & factories
98	TEMPLE	partial equivalence	protestant & churches
99	TEMPLE PAIEN	partial equivalence	ancient & temples
100	TENNIS	exact equivalence	tennis & courts (builtworks)
101	TONNELLERIE	exact equivalence	barrel (container) & factories
102	TREFILERIE	exact equivalence	wire & factories
103	USINE A GLACE	exact equivalence	ice & factories
104	USINE D'ACIDE SULFURIQUE	exact equivalence	sulfuric acid & factories
105	USINE D'ARMES	exact equivalence	ammunition & factories
106	USINE D'ARMES	partial equivalence	weapon & factories
107	USINE D'ARTICLES EN MATIERE PLASTIQUE	exact equivalence	plastic & hardware & factories
108	USINE D'EBENISTERIE	exact equivalence	cabinetmaking & factories
109	USINE D'ELEMENTS EN MATIERE PLASTIQUE POUR LE BATIMENT	exact equivalence	plastic & building material & factories
110	USINE D'ELEMENTS PREFABRIQUES	partial equivalence	prefabricated & building material & factories
111	USINE D'EMBALLAGE ET CONDITIONNEMENT	exact equivalence	packaging material & factories
112	USINE D'EMBALLAGES EN MATIERE PLASTIQUE	exact equivalence	plastic & packaging material & factories
113	USINE D'EMBOUTISSAGE	exact equivalence	stamping (forming) & factories
114	USINE D'ENCRES	exact equivalence	ink & factories
115	USINE D'ENGRAIS	exact equivalence	fertilizer & factories
116	USINE D'ESTAMPAGE	exact equivalence	cold & stamping (forming) & factories
117	USINE D'HORLOGERIE	exact equivalence	timepiece & factories
118	USINE D'IMPRESSION SUR ETOFFES	exact equivalence	cloth & printing & textile mills
119	USINE D'INSTRUMENTS DE MESURE	exact equivalence	measuring device & factories
120	USINE D'INSTRUMENTS DE MUSIQUE	exact equivalence	musical instrument & factories
121	USINE D'OUATE	exact equivalence	batting & factories
122	USINE D'OUVRAGES EN AMIANTE	exact equivalence	asbestos & factories
123	USINE DE BIMBELOTERIE	partial equivalence	wood & toy (recreational artifact) & factories
124	USINE DE BOISSELLERIE	partial equivalence	turning & factories
125	USINE DE BONNETERIE	exact equivalence	hosiery & factories
126	USINE DE BOUCHONS	exact equivalence	cork (bark) & factories
127	USINE DE BOUGIES	exact equivalence	candle & factories
128	USINE DE BOUTONS	exact equivalence	button (fastener) & factories
129	USINE DE BOYAUDERIE	exact equivalence	gut & factories
130	USINE DE BRODERIE MECANIQUE	exact equivalence	embroidering & factories
131	USINE DE CAOUTCHOUC	exact equivalence	rubber & factories
132	USINE DE CELLULOSE	exact equivalence	cellulose & factories
133	USINE DE CERAMIQUE	exact equivalence	ceramic & factories
134	USINE DE CHAPELLERIE	exact equivalence	hat & factories
135	USINE DE CHAUSSURES	exact equivalence	shoe (footwear) & factories
136	USINE DE CHAUX	exact equivalence	lime & factories
137	USINE DE COLLES	exact equivalence	glue & factories
138	USINE DE CONSTRUCTION AERONAUTIQUE	exact equivalence	aircraft & assembly plants
139	USINE DE CONSTRUCTION AUTOMOBILE	exact equivalence	automobile & assembly plants
140	USINE DE CONSTRUCTION METALLIQUE	exact equivalence	metal & building material & factories
141	USINE DE CONTRE PLAQUE	exact equivalence	plywood & factories
142	USINE DE COSMETIQUES	exact equivalence	cosmetic & factories
143,144	USINE DE CYCLES	exact equivalence	bicycle & factories **OR** motorcycle & factories
145	USINE DE DECOLLETAGE	partial equivalence	screw & factories
146	USINE DE DENTELLE MECANIQUE	exact equivalence	lace & factories
147	USINE DE DETERGENTS	exact equivalence	detergent & factories
148	USINE DE FABRICATION DE MATERIAUX DE CONSTRUCTION	exact equivalence	building material & factories
149	USINE DE FABRICATION ET DISTILLATION DES GOUDRONS	exact equivalence	tar & refineries

150	USINE DE FERBLANTERIE	exact equivalence	tinware & factories
151	USINE DE FEUTRE	exact equivalence	felt & factories
152	USINE DE FIBRE DE VERRE	exact equivalence	fiberglass & factories
153	USINE DE FIBRES ARTIFICIELLES ET SYNTHETIQUES	exact equivalence	synthetic fiber & factories
154	USINE DE FLACONNAGE	exact equivalence	bottle & factories
155	USINE DE GRES	exact equivalence	stoneware & factories
156	USINE DE MATERIEL AGRICOLE	exact equivalence	agricultural & equipment & factories
157	USINE DE MATERIEL DE TELECOMMUNICATION	exact equivalence	telecommunication & equipment & factories
158	USINE DE MATERIEL ELECTRIQUE INDUSTRIEL	partial equivalence	power producing equipment & factories
159	USINE DE MATERIEL FERROVIAIRE	exact equivalence	railroad car & assembly plants
160	USINE DE MATERIEL FERROVIAIRE	partial equivalence	locomotive & assembly plants
161	USINE DE MATERIEL INFORMATIQUE	exact equivalence	data processing & equipment & assembly plants
162	USINE DE MATERIEL MEDICOCHIRURGICAL	exact equivalence	medical & equipment & factories
163	USINE DE MATERIEL OPTIQUE	exact equivalence	optical instrument & factories
164	USINE DE MATERIEL PHOTO CINEMATOGRAPHIQUE	exact equivalence	photographic equipment & factories
165	USINE DE MATIERES COLORANTES SYNTHETIQUES	exact equivalence	synthetic dye & factories
166	USINE DE MATIERES PLASTIQUES	exact equivalence	plastic & factories
167	USINE DE MENUISERIE	exact equivalence	woodworking & factories
168	USINE DE MEUBLES	exact equivalence	furniture & factories
169	USINE DE PAPIERS PEINTS	exact equivalence	wallpaper & factories
170	USINE DE PARAPLUIES ET CANNES	exact equivalence	cane & factories
171	USINE DE PARAPLUIES ET CANNES	partial equivalence	umbrella & factories
172	USINE DE PASSEMENTERIE	partial equivalence	trimming & factories
173	USINE DE PEINTURES ET VERNIS	exact equivalence	varnish & factories
174	USINE DE PEINTURES ET VERNIS	partial equivalence	paint & factories
175	USINE DE PORCELAINE	exact equivalence	porcelain & factories
176	USINE DE POTERIE	partial equivalence	pottery & factories
177	USINE DE PRODUIT TEXTILE NON TISSE	partial equivalence	felting & factories
178	USINE DE PRODUITS CHIMIQUES	exact equivalence	chemical & factories
179	USINE DE PRODUITS EXPLOSIFS	exact equivalence	explosive & factories
180	USINE DE PRODUITS PHARMACEUTIQUES	exact equivalence	pharmaceutical & factories
181	USINE DE PRODUITS PHOTOGRAPHIQUES ET CINEMATOGRAPHIQUES	exact equivalence	photographic materials & factories
182	USINE DE QUINCAILLERIE	exact equivalence	metal & hardware & factories
183	USINE DE SERRURERIE	exact equivalence	lock (securing device) & factories
184	USINE DE SOUFRE	exact equivalence	sulfur & factories
185	USINE DE TABAC	exact equivalence	tobacco & factories
186	USINE DE TAILLE DE MATERIAUX DE CONSTRUCTION	partial equivalence	stonecutting & factories
187	USINE DE TAILLE DE PIERRE POUR LA JOAILLERIE ET L'INDUSTRIE	exact equivalence	lapidary & factories
188	USINE DE TRAITEMENT DE SURFACE DES METAUX	exact equivalence	plating & factories
189	USINE DE TRANSFORMATION DU LIEGE	exact equivalence	cork (bark) & factories
190	USINE DE VERRE CREUX	exact equivalence	glass & hollow-ware & factories
191	USINE DE VERRE PLAT	exact equivalence	plate glass & factories
192	USINE DE VERRES OPTIQUES	partial equivalence	optical glass & factories
193	USINE LIEE AU TRAVAIL DU BOIS	exact equivalence	woodworking & factories
194	VERRERIE	exact equivalence	glass & factories
195	VERRIERE EN COUVERTURE	exact equivalence	glass & roofs
196	VESTIAIRE D'USINE	exact equivalence	factory & locker rooms
197	VILLA	exact equivalence	ancient & villas
198	VINAIGRERIE	exact equivalence	vinegar & processing plants
199	VOIRIE	exact equivalence	city & streets