Augmenting thesaurus relationships: possibilities for retrieval

Douglas Tudhope[1], Harith Alani[2], Christopher Jones[3]

[1] **School of Computing, University of Glamorgan, Pontypridd, CF37 1DL, UK. dstudhope@glam.ac.uk**
[2] **Department of Electronics and Computer Science, Southampton University, SO17 1BJ, UK. ha@ecs.soton.ac.uk**
[3] **Department of Computer Science, Cardiff University, Cardiff, CF24 3XF, UK. C.B.Jones@cs.cf.ac.uk**

Abstract

This paper discusses issues concerning the augmentation of thesaurus relationships, in light of new application possibilities for retrieval. We first discuss a case study that explored the retrieval potential of an augmented set of thesaurus relationships by specialising standard relationships into richer subtypes, in particular hierarchical geographical containment and the associative relationship. We then locate this work in a broader context by reviewing various attempts to build taxonomies of thesaurus relationships and conclude by discussing the feasibility of hierarchically augmenting the core set of thesaurus relationships, particularly the associative relationship. We discuss the possibility of enriching the specification and semantics of RT relationships, while maintaining compatibility with traditional thesauri via a limited hierarchical extension of the associative (and hierarchical) relationships. This would be facilitated by distinguishing the type of term from the (sub)type of relationship and explicitly specifying semantic categories for terms following a faceted approach.

We first illustrate how hierarchical spatial relationships can be used to provide more flexible retrieval for queries incorporating place names in applications employing online gazetteers and geographical thesauri. We then employ a set of experimental scenarios to investigate key issues affecting use of the associative (RT) thesaurus relationships in semantic distance measures. Previous work has noted the potential of RTs in thesaurus search aids but also the problem of uncontrolled expansion of result sets. Results presented in this paper suggest a potential for taking account of the hierarchical context of an RT link and specialisations of the RT relationship.

1. Introduction

Recent years have seen convergence of work in digital libraries, museums and archives with a view to resource discovery and opening up access to digital collections. Various projects are following standards-based approaches building upon terminology and knowledge organisation systems (Hodge 2000). Concurrently, within the web community, there has been growing interest in vocabulary-based techniques, with the realisation of the challenges posed by web searching and retrieval applications. This has manifested itself in metadata initiatives, such as Dublin Core and the proposed W3C Resource Description Framework. In order to support retrieval, provision is made in such metadata element sets for thematic keywords from vocabulary tools such as thesauri (ISO 2788, ISO 5964). Ontologies incorporating thesauri or related semantic models underpin diverse ongoing projects in remote access, quality-based services, cross domain searching, semantic interoperability, building RDF models and digital libraries generally (Amann and Fundulaki 1999; Doerr and Fundulaki 1998; Koch 2000; Michard and Pham-Dac 1998).

This paper is in two parts. We first discuss a case study that explored the retrieval potential of an augmented set of thesaurus relationships by specialising standard relationships into richer subtypes, in particular hierarchical geographical containment and the associative relationship. We then locate this work in a broader context by reviewing various attempts to build taxonomies of thesaurus relationships and conclude by discussing the feasibility of hierarchically augmenting the core set of thesaurus relationships, particularly the associative relationship. The work described here was part of a larger project, OASIS (Ontologically Augmented Spatial Information System), exploring terminology systems for thematic and spatial access in digital library applications. One of its aims concerned the retrieval potential of spatial metadata with rich place name data but limited locational data (footprint). Such representations occur in online gazetteers, geographical thesauri or geographic name servers, when conventional GIS datasets are unavailable, unnecessary or pose undesirable bandwidth limitations (Hill 2000; Jones 1997).

Another aim was to explore the potential of reasoning over the semantic relationships in thesauri to assist retrieval. The three main thesaurus relationships are Equivalence (equivalent terms), Hierarchical (broader/narrower terms: BT/NTs), Associative (Related Terms: RTs). Studies support the use of thesauri in online retrieval and the potential for combining free text and controlled vocabulary approaches (e.g. Fidel 1991). However there are various research challenges, including the 'vocabulary problem' – differences in choice of index term at different times by indexers and searchers (Chen *et al.* 1997). Indexer and searcher may be operating at different levels of specificity, and at different times an indexer(s) may make different choices from a set of possible term options. While conventional narrower term expansion may help in some situations, a more systematic approach to thesaurus term expansion has the potential to improve recall in such situations. In the work described here, we have employed the Getty AAT and TGN vocabularies[1]. Harpring (1999) gives an overview of the Getty's vocabularies with examples of their use in web retrieval interfaces and collection management systems. Examples are given of their use as a source of variant names of a concept. It is suggested that the AAT's RT relationships may be helpful to a user exploring topics around an information need and the issue of how to perform query expansion without generating too large a result set is also raised.

In section 2 we discuss our schema, illustrating how the spatial relationships in the thesaurus can be used to provide more flexible retrieval for queries incorporating place names. The second topic (sections 3 and 4) concerns the use of associative thesaurus relationships in retrieval. Existing collection management systems include access to thesauri for cataloguing with fairly rudimentary use of thesauri in retrieval (mostly limited to interactive query expansion/refinement and Narrower Term expansion). In particular, there is scope for increased use of associative (RT) relationships in thesaurus-based retrieval tools. There is a danger that incorporating RTs into retrieval tools with automatic query expansion may lead to uncontrolled expansion of result sets. We discuss experimental scenarios involving semantic distance measures in order to map key issues affecting use of RTs. Section 5 reviews various taxonomies of augmented thesaurus relationships while Section 6 discusses the potential for a limited extension of the core set of relationships. Conclusions are outlined in section 7.

2. OASIS Overview and Spatial Access Example

We adopted collection data from the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) database of Scottish archaeological sites and historical buildings (Murray, 1997). The AAT provided thematic index terms such as 'arrow', 'bronze', 'axe', 'castle', etc. The spatial data in the OASIS system includes information on hierarchical and adjacency relations between named places, in addition to place types, and (centroid) co-ordinates. This information was taken from the TGN (Harpring 1997), augmented with data derived from the Bartholomew's (Harper Collins 2000) digital map data for Scotland.

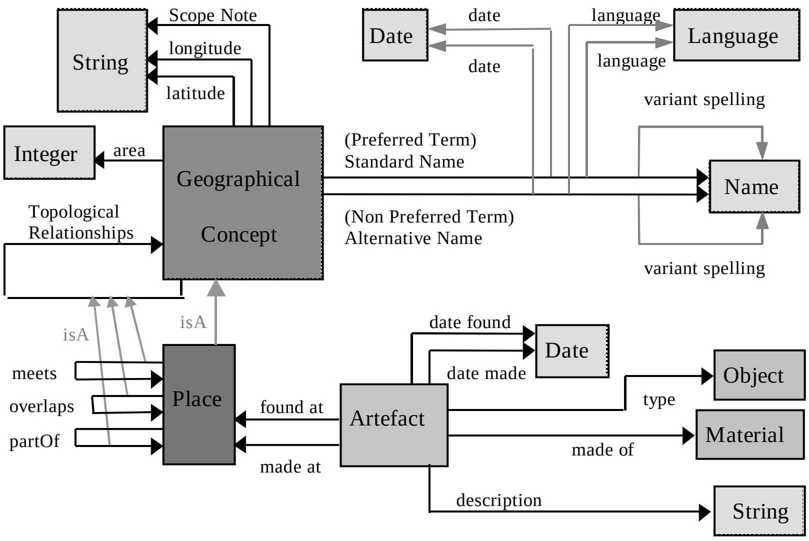

Figure 1. OASIS schema for *Place* and *Artefact*

The term 'ontology' has widely differing uses in different domains (Guarino 1995). Our usage here follows that of Amann and Fundulaki (1999), in that we see an ontology as a conceptualisation of a domain, in effect providing a connecting semantics between thesaurus hierarchies with specifications of roles for combining thesaurus elements. The OASIS schema (Figure 1) encompasses different versions of place names (e.g. current and historical names, different spellings, etc.), place types (e.g. Town, Building, Port, River, Hill), latitude and longitude co-ordinates, and topological relationships (e.g. meets, part of). The schema is implemented using the object-oriented Semantic

[1] Art and Architecture Thesaurus; Thesaurus of Geographic Names

Index System (SIS - Constantopolous and Doerr 1993) also used to store the data, and which provided the AAT implementation. The SIS has a meta modelling capability and an application interface for querying the schema. Figure 1 shows the meta level classification of the classes *Place* and *Artefact*. As we discuss later in relation with RTs, relationships can be instantiated or subclassed from other relationships. Thus, *meets*, *overlaps*, and *partOf* are subclasses of *Topological Relationships*. The information stored in the OASIS database can be accessed using a set of functions through which it is possible to find all information related to a given place, or to find all places with a given spatial relationship, or to find objects at a place made of a certain material. For example to find all the places that are part of the City of Edinburgh, the system would return a set of all the places that are linked with a geographical *partOf* relationship to the City of Edinburgh.

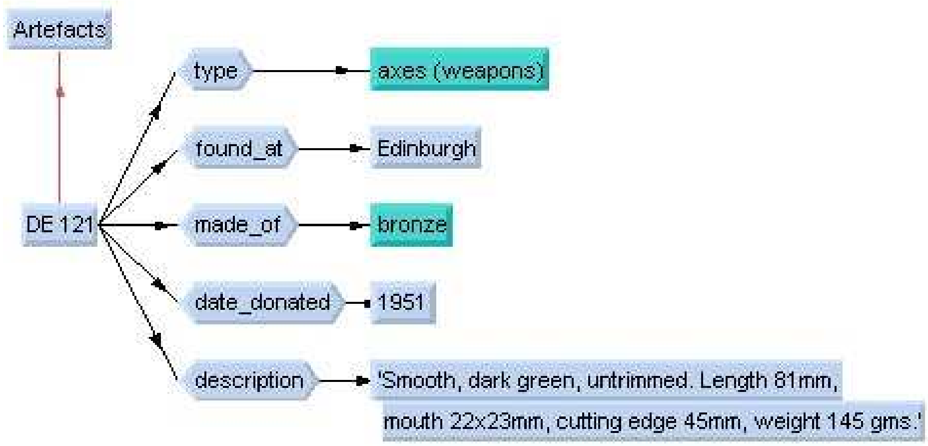

Figure 2. Classification of the axe artefact NMRS Acc. No. DE 121

Figure 2 shows the OASIS schema for a particular object (axes are a common type of archaeological artefact in the RCAHMS dataset). OASIS implements a set of thematic and spatial measures that enables query expansion to find similar terms. Conventional GIS measures (eg zone buffering) could be applied in situations where a full GIS polygon dataset is available. However, as discussed earlier, there are situations where a GIS is not available or unnecessary. Consider the query *Do you have any information on axes found in the vicinity of Leith?* An exact match to the query would only return axes indexed by the term *Leith* (a district of the city of Edinburgh). To search for axes found in the vicinity of Leith, spatial distance measures can be applied to expand the geographical term *Leith* to spatially similar places, where axes have been found. These places can be ranked by spatial similarity using the *part-of* spatial containment relationship, which in OASIS is based on the spatial hierarchies in the TGN. As we discuss in Section 5, this relationship is a subtype of the hierarchical thesaurus relationship. Given the term *Leith*, the OASIS spatial hierarchy distance measure would produce the list of places in Table 1, ranked according to their spatial hierarchical similarity to *Leith*. Some places such as *Corstorphine*, *Edinburgh*, *Currie* score highly, since (like *Leith)* they are districts within the region *City of Edinburgh*. Similarly, since *Penicuik*, *Broxburn*, *Inveresk*, etc are places in Scotland, they would be returned ahead of any axe finds in England.

Place	Score	Place	Score
Edinburgh‘Leith	100	Midlothian‘Penicuik	35
Edinburgh‘Edinburgh	60	Midlothian‘Temple	35
Edinurgh‘Corstorphine	60	West_Lothian‘Kirknewton	35
Edinburgh‘Currie	60	East_Lothian‘Pencaitland	35
Edinburgh‘Duddingston	60	West_Lothian‘Broxburn	35
Edinburgh‘Dalmeny	60	Midlothian‘Leadburn	35
Edinburgh‘Ratho	60	Midlothian‘Fala	35
Edinburgh‘Kirkliston	60	West_Lothian‘Mid_Calder	35
East_Lothian‘Musselburgh	35	East_Lothian‘East_Saltoun	35
East_Lothian‘Inveresk	35	East_Lothian‘Bolton	35
Midlothian‘Dalkeith	35	West_Lothian‘Livingston	35
Midlothian‘Borthwick	35		

Table 1. A list of places ranked according to their similarity to Leith using the Spatial Hierarchical measure

The TGN also provides centroid co-ordinate data for places/regions – used by OASIS in a Euclidean distance measure. Table 2 shows the places of the previous table, now ranked according to their similarity to *Leith* using an integration of the spatial hierarchical and Euclidean distance measures. In some situations (e.g. queries relating to administrative responsibilities), administrative hierarchies may be quite relevant but not an overriding factor in judgements of spatial similarity and thus we provided a combination of the two measures. It can be seen in Table 2 that the rankings now also take account of Euclidean distance (for example *Musselburgh* compared to *Livingstone).*

Place	Score	Place	Score
Edinburgh‘Leith	100	Midlothian‘Penicuik	69
Edinburgh‘Duddingston	89	Midlothian‘Temple	69
Edinburgh‘Edinburgh	88	West_Lothian‘Kirknewton	67
Edinurgh‘Corstorphine	83	East_Lothian‘Pencaitland	66
Edinburgh‘Currie	81	West_Lothian‘Broxburn	66
Edinburgh‘Dalmeny	78	Midlothian‘Leadburn	66
East_Lothian‘Musselburgh	77	Midlothian‘Fala	65
Edinburgh‘Ratho	77	West_Lothian‘Mid_Calder	64
East_Lothian‘Inveresk	76	East_Lothian‘East_Saltoun	63
Edinburgh‘Kirkliston	76	East_Lothian‘Bolton	62
Midlothian‘Dalkeith	73	West_Lothian‘Livingston	60
Midlothian‘Borthwick	71		

Table 2. A list of places ranked using the Spatial Hierarchical and Euclidean measures

Euclidean distance between centroid coordinates is less satisfactory for large area regions. Voronoi-based techniques can be employed with limited centroid footprint metadata to yield a richer approximation of spatial regions. Our larger project has investigated this boundary approximation method, combining it with geographical thesaurus relationships (Alani *et al*. 2001). The method has the potential to assist a range of spatial queries.

3. The retrieval potential of the associative relationship

A thesaurus can act as a search aid by providing a set of controlled terms that can be browsed via some form of hypertext representation (e.g. Bruza 1990; Pollitt 1997). This can assist the user to understand the context of a concept, how it is used in a particular thesaurus and provide feedback on number of postings for terms (or combinations of terms). The inclusion of semantic relationships in the index space, moreover, provides the opportunity for knowledge-based approaches where the system takes a more active role in building a query by automatic reasoning over the relationships (Cunliffe *et al.* 1997; Tudhope and Cunliffe 1999). Candidate terms can be suggested for a user to consider in refining a query and various forms of query expansion are possible. For example, items indexed by terms semantically close to query terms can be included in a ranked result list and imprecise matching between two media items is useful in ‘More like this’ options. The basis for such automatic term expansion is some kind of semantic distance measure, often based on the minimum number of semantic relationships that must be traversed in order to connect the terms (Rada *et al* 1989). For a review of semantic distance measures and weighting factors that have been employed, see Alani *et al.* (2000).

RTs represent a class of non-hierarchical relationships, which have been less clearly understood in thesaurus construction and applicability to retrieval than the hierarchical relationships. At one extreme, an RT is sometimes taken to represent nothing more than an extremely vague ‘See-also’ connection between two concepts. This can lead to uncontrolled expansion of result sets when RT relationships are expanded and a potential loss in precision. Rada *et al.* (1989) argue that semantic distance measures over RT relationships can be less reliable than over hierarchical relationships, unless the user's query can be closely linked to the RT relationship. The basic assumption of a cognitive basis for a semantic distance effect over thesaurus terms has been investigated by Brooks (1997), in a series of experiments exploring the relevance relationships between bibliographic records and topical subject descriptors. These studies, employing the ERIC database and linked thesaurus, involved strictly linear hierarchies, as opposed to tree hierarchical structures (as with the AAT) or indeed poly-hierarchies. However the results are suggestive of the existence of some semantic distance effect, with an inverse correlation between semantic distance and relevance assessment, dependant on position in the subject hierarchy, direction of term traversal and other factors. In particular, a definite effect was observed for RTs (typically less than for hierarchical traversal). An empirical study by Kristensen (1993) compared single-step automatic query expansion of synonym, narrower-term, related term, combined union expansion and no expansion of thesaurus relationships. Thesaurus expansion was found to improve recall significantly at some (lesser) cost in precision. Taken separately, single step RT expansion results did not differ significantly from NT or synonym expansion. In another empirical study (Jones *et al.* 1995), a log was kept of users' choices of relationships interactively expanded via thesaurus navigation while entering a query. In this study of users refining a query, a majority of terms retrieved from the thesaurus came from RTs (the then INSPEC thesaurus contained many more RTs than hierarchical relationships).

4. Case study of RT retrieval scenarios

This section maps key issues affecting use of RTs in term expansion algorithms for retrieval. Results are given from a series of scenarios applying different versions of a semantic distance algorithm to terms in the AAT (AAT 2000). The distance measure employed a branch and bound algorithm, with a depth factor which reduced costs according to hierarchical depth (Tudhope and Taylor 1997). It was implemented in C++ using the SIS function library to query the underlying schema given in Figure 1. For the purposes of the scenarios, the threshold used to terminate expansion was 2.33.

Our aim was to investigate different factors relevant to RT expansion, rather than relative weighting of relationships. The weights for this experiment were selected to reflect some broad consensus of previous research (see Alani *et al.* 2000). Our weights (BT 3, NT 3, RT 4), taken together with a depth factor inversely proportional to the hierarchical depth of the destination term, assign lowest costs to NTs and favour RTs over BTs at higher depths in the hierarchy (following an AAT editorial observation that RTs appear to work better at fairly broad levels).

We developed a series of experimental scenarios based around term generalisation involving RT traversal. Building on the example in Section 2, we first focus on the AAT's *Objects* Facet: *Weapons & Ammunition* and *Tools & Equipment* hierarchies. The initial scenario supposes a narrowly defined information need for items concerning axes used as weapons (mapping to AAT term *axes (weapons)*). In this initial scenario, expansion is limited and restricted to NT relationships only (shown in plain black text in Table 3): *tomahawks*, *battle-axes*, *throwing axes*, and *franciscas*.

The second scenario supposes an information need for items more broadly connected with axes used as weapons thus allowing for some flexibility in expansion. We first consider expansion only over hierarchical relationships and then discuss expansion with RTs. Table 3 shows results from BT/NT expansion, with semantic distance shown for each term (terms in green italics result from expansion over both BT and NT relationships as opposed to strict NT expansion downwards from *axes (weapons)*).

Term	Distance
axes (weapons)	0
tomahawks	0.6
battle-axes	0.6
edged weapons	*1*
throwing-axes	1.1
franciscas	1.53
staff weapons	*1.75*
sword sticks	*1.75*
harpoons	*1.75*
bayonets	*1.75*
daggers (weapons)	*1.75*
fist-weapons	*1.75*
knives (weapons)	*1.75*
swords	*1.75*

Table 3. BT/NT expansion only

Table 4 shows the effect of introducing RT expansion (new terms in blue italics). Staff weapons relating to axes *(halberds, pollaxes, gisarmes)* move up the ranking and are now below the threshold. Other new terms (such as *axes (tools), chip axes, ceremonial axes)* are also introduced. These terms could well be relevant to broader information needs or to situations when the initial thesaurus entry term was mismatched (for example, when an information need related more to tool use than weapons). In some situations however, the new terms could be seen as ‘noise’ and finer grained control of RT expansion would be desirable.

Term	Distance	Term	Distance
axes (weapons)	0	bayonets	1.75
tomahawks (weapons)	0.6	daggers (weapons)	1.75
battle-axes	0.6	fist weapons	1.75
edged weapons	1	swords	1.75
axes (tools)	*1*	*<projectiles with nonexplosive propellant>*	*1.77*
halberds	*1*	*adze-hatchets*	*1.9*
pollaxes	*1*	*hewing hatchets*	*1.9*
gisarmes	*1*	*lathing hatchets*	*1.9*
ceremonial axes	*1*	*shingling hatchets*	*1.9*
throwing axes	1.1	*<cutting tools>*	*2*
hatchets	*1.4*	*fasces*	*2*
franciscas	1.53	*Pulaskis*	*2*
chip axes	*1.6*	*<ceremonial weapons>*	*2*
berdyshes	*1.6*	*<wood-cutting and - finishing tools>*	*2.15*
staff weapons	1.75	*arrows*	*2.33*
sword sticks	1.75	*machetes*	*2.33*
harpoons	1.75	*darts*	*2.33*

Table 4. RT and BT/NT expansion - terms in blue italics are introduced with RT expansion

As seen in section 5, the ISO standard and other reviews of RTs in thesaurus practice make a distinction between RTs within the same hierarchy and RTs between hierarchies (or sometimes facets). One method of achieving finer control in RT expansion is to filter on the original term's (sub)hierarchy - RTs to terms within different sub-hierarchies are not traversed. Table 5 shows the comparison with the previous example in Table 4. Terms in red underline (mostly from the Tools&Equipment hierarchy) would now be excluded. Thus the effects of RT expansion within the hierarchy have been retained while the number of additional terms has been reduced. Other options are possible for semantic distance measures and term expansion in general. If information on facets and hierarchies was retained in thesaurus database implementations, it would be possible to weight RT traversal differentially according to the hierarchies/facets linked. On the negative side, note that in this scenario axes that are both tools and weapons *(hatchets, machetes)* are excluded, since due to the monohierarchical representation of the AAT[2] they are located within the *Tools&Equipment* hierarchy. In many situations this will not be desirable.

Term	Distance	Term	Distance
axes (weapons)	0	bayonets	1.75
tomahawks (weapons)	0.6	daggers (weapons)	1.75
battle-axes	0.6	fist weapons	1.75
edged weapons	1	swords	1.75
axes (tools)	1	*<projectiles with nonexplosive propellant>*	*1.77*
halberds	*1*	adze-hatchets	1.9
pollaxes	*1*	hewing hatchets	1.9
gisarmes	*1*	lathing hatchets	1.9
ceremonial axes	*1*	shingling hatchets	1.9
throwing axes	1.1	<cutting tools>	2
hatchets	1.4	fasces	2
franciscas	1.53	Pulaskis	2
chip axes	1.6	*<ceremonial weapons>*	*2*
berdyshes	*1.6*	<wood-cutting and - finishing tools>	2.15
staff weapons	1.75	*arrows*	*2.33*
sword sticks	1.75	machetes	2.33
harpoons	1.75	*darts*	*2.33*

Table 5. RT expansion – red underlined terms excluded when inter-hierarchical RT traversals are not allowed

The next scenario explores an alternative approach to control of RT expansion based upon selective specialisation of the associative relationship according to retrieval context[3]. The aim is to take advantage of more structured approaches to thesaurus construction where different types of RTs are employed. In some circumstances it may be appropriate to consider all types of associative relationships as a generic RT for retrieval purposes (as in the above scenarios). However, under other contexts it may be desirable to treat RT sub-types differently, permitting some RT traversals but forbidding or penalising (via weighting) others. Thus heuristics may selectively guide RT expansion, depending on query model and session context. The AAT is particularly suited to investigation of this topic, since its editors followed a systematic, rule-based approach to the design of RT links (Molholt 1996). The AAT RT editorial manual specifies a set of rules to apply to the relevant hierarchical context and scope notes in order to identify valid RT relationships between terms when building the vocabulary or enhancing it. This includes a set of specialisations of the RT relationships (AAT 1995; and see also extract in Table 6), following their notation: *1A and 1B)* Alternate hierarchical (BT/NT) relationships (since AAT is not polyhierarchical); *2A and 2B)* Part/Whole relationships; *3)* Several Inter/intra Facet relationships (eg Agents-Activities and Agents-Materials); *4)* Distinguished From relationship (the scope note evidences a need to distinguish the sense of two terms); *5)* frequently Conjuncted terms (eg Cups AND Saucers). We have extended the original SIS AAT schema to

[2] The AAT is conceptually polyhierarchical but is currently physically monohierarchical due to the original database software employed in the project. There are plans to port it into a polyhierarchical data structure.
[3] This is in keeping with the recommendation of Rada *et al.* (1991) that automatic expansion of non-hierarchical relationships be restricted to situations where the type of relationship can be linked with the particular query, and also with Jones' (1993) discussion of using sub-classifications to help distinguish relationships according to strength.

specialise the associative relationship (see Alani *et al.* 2000). RTs in our schema can optionally be treated as specialised sub-relationships, or as generic RTs.

1. Alternate BT (1A) and NT (1B)
--- allows a polyhierarchy to be represented in a monohierarchical software system
Eg) <telecommunication structures>*radio towers* RT-1A <single built works by form>*towers*

2. Whole/Part (2A) and Part/Whole (2B)
Note this relationship may or may not be reciprocal (eg, when a whole may or may not have a part but when the part exists it is always part of the whole).
Eg) *stalls* RT-2B *barns*
Eg) *fortifications* RT-2A *fortification elements* (the two terms belong to different hierarchies)

3. Code 3 represents various RT links between hierarchies and facets. These include:

3A Concept/User or Creator

3B Concept/Resulting or causative Action

3H Agents/Objects<hierarchies relating to location or setting>

3J Agents/Objects<hierarchies relating to furnishings or equipment used by these agents>

3K Agents/Objects<h. relating to products used or created by these agents>

3S Materials/Objects where a strong link exists between a Materials facet term and an Objects term

3T Objects<h. relating to furnishing/equipment>/Objects/<h. relating to its location or setting>

3U Objects<h. relating to visual/verbal work>/Objects/<h. relating to its location or setting>

3W Objects<h. relating to furnishing/equipment>/Objects/<h. relating to related visual/verbal work>

4. RT link between terms mentioned or implied as ‘distinguished from’ in Scope Notes.
Eg) *quadrangles* RT-4 *courtyards*

5. RT link between two terms typically conjoined by AND but which ended up as separates
Eg) *cups* RT-5 *saucers*

Table 6. Extract from AAT Related Term Guidelines
For a full definition, see AAT(1995).

The editorial rules for creating specific associative relationships are not retained in electronic implementations of the AAT to date. Therefore, for this experiment we manually specialised all RT relationships 3 links away from *axes (weapons)* into their corresponding sub-types by following sample extracts of AAT Editorial Related Term Sheets and applying the editorial rules. Figure 3 shows the resulting visualisation of the concept *axes (weapons)* after specialising the RT relationships – note the two different subtypes of RT. In the next scenario, the distance algorithm was set to filter on the subtype of RT, only permitting traversal over the Alternate BT and Alternate NT relationships. Table 7 shows the results (terms now excluded from Table 4 shown in red underline). The effect can be compared with terms excluded by the hierarchy filtering approach in Table 5. This scenario might correspond to a reasonably strict information request but where some terms located in the *Tools & Equipment* hierarchy were relevant. For example, an alternate NT relationship exists between *tomahawks* and *hatchets*. Since they are classed as both tools and weapons, *hatchets* might well be regarded as relevant to the scenario. Terms, such as *machetes* and *hatchets* from the *Tools & Equipment* hierarchy were excluded when narrowly filtering on the hierarchy but are now included. The specialisation permits the AAT to be treated as polyhierarchical for retrieval.

Figure 3, an extract from the AAT's *Tools&Equipment* and *Weapons&Ammunition* hierarchies, focuses on the RT relationships connecting the hierarchies[4] – (see AAT 2000 for a display of the full hierarchies). Two different types of RT are represented. The AAT Scope Note for *axes (weapons)* reads:

> *"Cutting weapons consisting basically of a relatively heavy, flat blade fixed to a handle, wielded by either striking or throwing. For axes used for other purposes, typically having narrower blades, use axes (tools)."*

Thus the associative relationship between *axes (weapons)* and *axes (tools)* is of subtype *Distinguished From* (see Table 6) and is not traversed in this scenario when filtering only on alternate hierarchical RT subtypes. We can see in Table 7 that the term *axes (tools)* and tool-related terms derived solely from this link *(chip axes, cutting tools, etc)* are excluded. Under some contexts, such terms might be considered of relevance but in a stricter weapons-related scenario they might well be seen as less relevant and can now be suppressed. The point is that this control can be passed to the retrieval system.

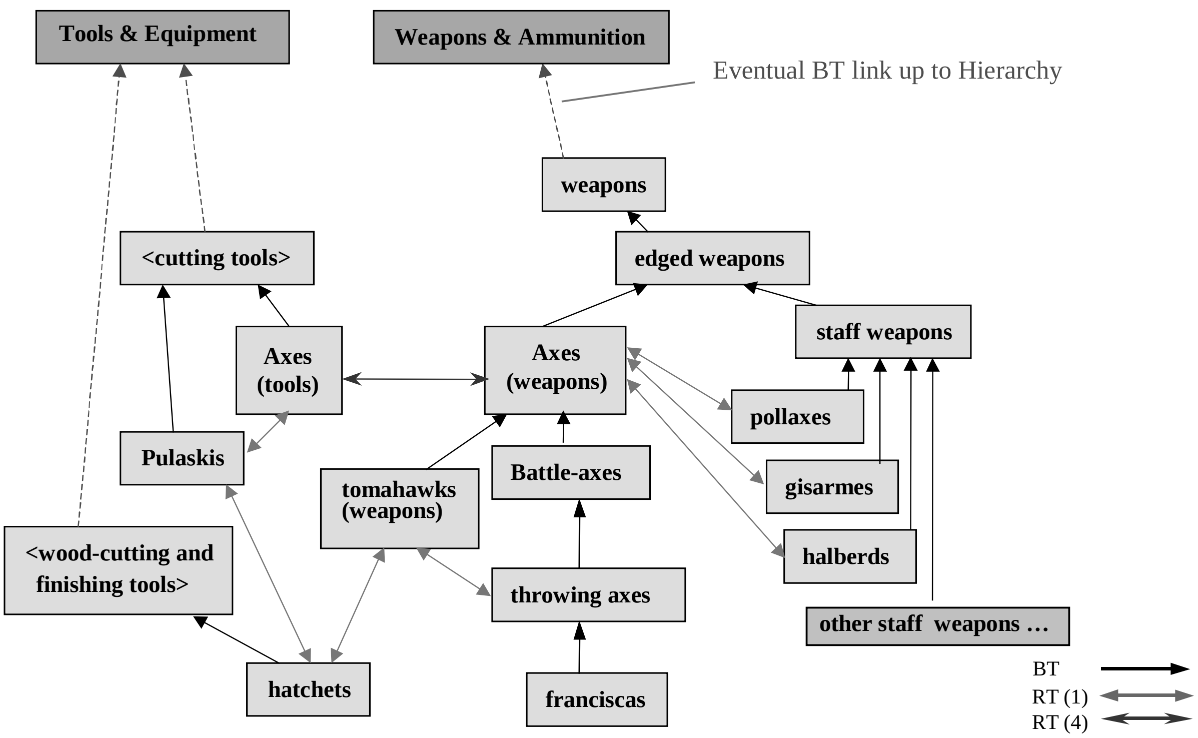

Figure 3. Extract of AAT around *axes (weapons)* with specialised RT relationships

[4] Note that staff weapons not connected by an RT to *Axes (weapons)* have been omitted due to space restrictions.

Term	Distance	Term	Distance
axes (weapons)	0	bayonets	1.75
tomahawks (weapons)	0.6	daggers (weapons)	1.75
battle-axes	0.6	fist weapons	1.75
edged weapons	1	swords	1.75
axes (tools)	1	*<projectiles with nonexplosive propellant>*	*1.77*
halberds	*1*	*adze-hatchets*	*1.9*
pollaxes	*1*	*hewing hatchets*	*1.9*
gisarmes	*1*	*lathing hatchets*	*1.9*
ceremonial axes	*1*	*shingling hatchets*	*1.9*
throwing axes	1.1	<cutting tools>	2
hatchets	*1.4*	fasces	2
franciscas	1.53	*Pulaskis*	*2*
chip axes	1.6	*<ceremonial weapons>*	*2*
berdyshes	*1.6*	*<wood-cutting and - finishing tools>*	*2.15*
staff weapons	1.75	*arrows*	*2.33*
sword sticks	1.75	*machetes*	*2.33*
harpoons	1.75	*darts*	*2.33*

Table 7. Filtering by RT subtype (alternate hierarchical RTs only) – red underlined terms now excluded from Table 4

Other scenarios illustrate the potential for filtering on other types of RT relationship. For example, an information need relating to *archery and its equipment*, might justify traversal of AAT RT inter-facet subtype *Activity - Equipment Needed or Produced*. This would yield the terms *arrows* and *bows (weapons)*, which could in turn be expanded to terms such as *bolts (arrows)*, *crossbows*, *composite bows*, *longbows*, and *self bows*. The same approach can be applied to scenarios relating to parts or components of an object, using the RT *Whole/Part*, and *Part/Whole* subtypes. Thus, a query on *arrows* (Figure 4) yields the terms listed in Table 8, using an expansion threshold of 1.3. Again, for this scenario we manually specialised all RT relationships 3 links away from *arrows* into their corresponding sub-types by following sample extracts of AAT Editorial Related Term Sheets and applying the editorial rules. The terms retrieved through Alternate Hierarchical RTs and Whole/Part RTs are shown in blue italics and green italics (arial font) respectively. For example, note that subparts *feathers* and *arrowheads* are included in the results.

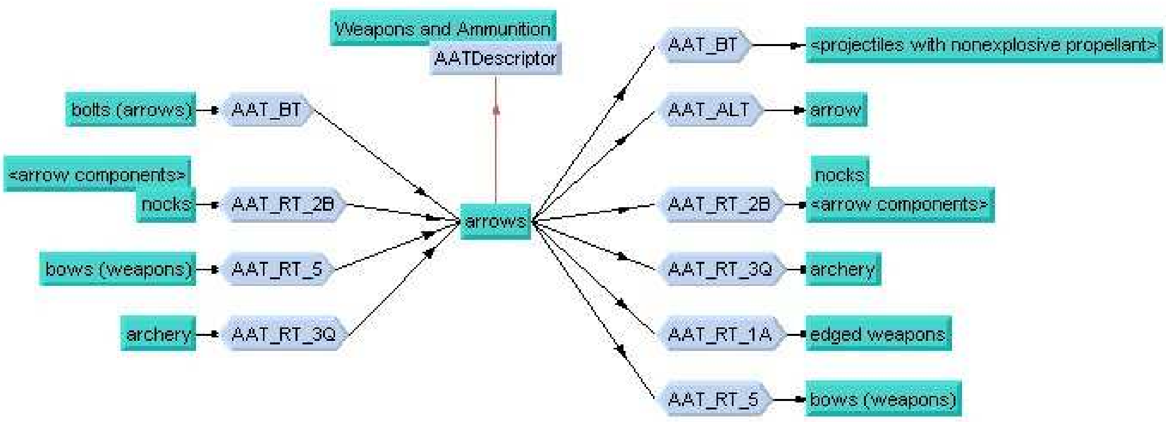

Figure 4. AAT visualisation of *arrows* with RT specialisations

Term	Distance	Term	Distance
Arrows	0	*harpoons*	*1.09*
bolts (arrows)	0.43	*bayonets*	*1.09*
<projectiles with nonexplosive propellant>	0.60	*daggers (weapons)*	*1.09*
<arrow components>	*0.67*	*fist weapons*	*1.09*
Nocks	*0.67*	*knives (weapons)*	*1.09*
Edged weapons	*0.67*	*swords*	*1.09*
Crossbows	*1.00*	darts	1.10
Arrowheads	*1.09*	throwing knives	1.10
Feathers (arrow components)	*1.09*	throwing-sticks (projectiles)	1.10
staff weapons	*1.09*	bolas	1.10
axes (weapons)	*1.09*	*<projectile weapons components>*	*1.27*
Sword sticks	*1.09*	*weapons*	*1.27*

Table 8. RT expansion: filtering on Alternate Broader/Narrower and Whole/Part subtypes

5. Review of thesaurus relationship taxonomies

Semantic modelling occurs in various computing domains. The standards for thesauri and related knowledge organisation systems in information science can be distinguished from the semantic structures common in AI or database modelling (eg Brachman 1983, Storey 1993[5]) by a particular emphasis on retrieval and interoperability across different subject domains. A well established information science tradition[6] allows software for collection management, thesaurus representation and retrieval applications to be shared across thesauri in different domains. The tradition rests on the core set of thesaurus relationships (equivalence, hierarchical and associative) mentioned earlier. The disciplining of semantic relationships to this core set makes possible the various aspects of interoperability by providing a stable, manageable foundation for different types of application.

Traditional use has tended to rely on human inspection of thesaurus representations, for example interactively browsing term hierarchies or manually looking up thesaurus displays in print form. Recently, with the growth of online collections, we have seen a move to enhanced machine processing of thesaurus representations and this has motivated a concern with extending the core set of relationships. Examples of these new applications include investigations of query expansion techniques in retrieval (Beaulieu 1997, Tudhope and Taylor 1997), efforts to devise automated mapping between different thesauri for cross domain or multi-lingual searching (Doerr 2000) and proposals for RDF representations of thesauri for the emerging 'semantic web' (Amann and Fundulaki 1999; Cross *et al.* 2000). Human interpretation of the context to infer the particular instance of a relationship type or tacit rules underlying facet structure are no longer a resource in automated traversal of the semantic network. Support for this trend towards an augmented set of relationships can be found in the ALA Subject Analysis Committee Final Report (ALA 1999), which supported a richer set of relationships, expressed in hierarchies, and also in the NISO Report on the Workshop on Electronic Thesauri (Milstead 1999) which advocated a "core set of relationships, hierarchically organized" (but not any minimal set). A richer set of relationships would assist efforts in these new application areas for thesauri by allowing finer grained automated reasoning. It would also converge with work on broader ontological conceptualisations, attempting to more formally define the roles played by entities in the schema (e.g. Bechofer 2000; Bechofer and Goble 1999). However, there is a danger that an undisciplined expansion of the underlying semantic model would lose the battle for interoperability. There is also a need for compatibility with the large number of existing thesauri (the Association for Information Management has a library of over 600 thesauri).

[5] Storey (1993) reviews the use of semantic relationships in data modelling and automated database design tools. She discusses a taxonomy of seven types of relationships, including various partitive (meronymic) relationships, and derives guidelines for representing them in a relational data model.
[6] The social aspect is also important; educational material and training practices distribute techniques for cataloguing and searching widely and promote good practice. Standards must also facilitate the modification of knowledge organisation systems, as the corresponding field of study evolves.

The following discussion connects the case study discussed in the first half of the paper with possibilities for limited extensions to the standard set of relationships, in particular focusing on the problems inherent in attempting to extend the associative relationship.

5.1 Hierarchical relationships

We first briefly consider hierarchical thesaurus relationships. There are three commonly accepted subtypes of the hierarchical relationship (ISO 2788), which might form a natural second level of hierarchical relationships for consideration in any standard extension of thesaurus relationships:

1. Generic (subclass/superclass)
2. Instance (class/instance)
3. Whole-Part (partitive): This is a hierarchical relationship between concepts of the same type, where 'the name of the part implies the name of the possessing whole in any context'. ISO 2788 allows four partitive cases:
    - 3a Systems and organs of the body
    - 3b Geographical location (as in our spatial query expansion examples in Section 2)
    - 3c Discipline (or field of study)
    - 3d Social structures

5.2 Associative relationships

Associative relationships are more difficult to specify. The notion of distinguishing subtypes of RTs has surfaced from time to time, usually with respect to proposed editorial methods for asserting associative relationships between terms. The 1986 ISO Guide to establishment and development of monolingual thesauri (ISO 2788) gives examples of several subtypes of the associative relationship, although these are intended to be representative examples from practice rather than any definitive list. The Standard suggests that frequently one RT term will occur in any definition of the other (eg in a Scope Note). To prevent precision being unnecessarily degraded by RTs unlikely to be of practical use, the Standard recommends that a term linked by an associative relationship be strongly implied by the other "according to the frames of reference shared by users of the index". Thus RT practice can vary according to the intended purpose of the thesaurus. The standard first identifies two types of term that can be linked by an associative relationship: those belonging to the same 'category' and those bridging categories. In practice, category is usually taken to be a thesaurus hierarchy and thus the distinction is essentially *intra* versus *inter hierarchy* (as in the discussion around Table 5 in Section 4). RTs should not usually occur between sibling terms, since there is a strong hierarchical connection between two siblings. (However, it can be appropriate to make an RT between two siblings when there exists a particularly strong relationship between them which does not extend to the other siblings - the AAT RT subtype *Distinguished From* often relates siblings). Without intending to be exhaustive, the Standard goes on to list typical examples of inter-hierarchical RTs:

ISO1 Intra-hierarchical
ISO2 Inter-hierarchical
ISO2a Discipline – Object of study
ISO2b Operation/Process – Agent/instrument
ISO2c Action – Product of action
ISO2d Action – Patient of action
ISO2e Concept - Property
ISO2f Concept – Origin
ISO2g Causal
ISO2h Thing – Counter agent
ISO2i Concept – Unit of measurement
ISO2j Phrases – Embedded nouns

Aitchison & Gilchrist (1987, p44*)* suggest some other RT subtypes in their influential explication of the Standard:

Part - Whole[7]
Occupation – Agent/Person
Material – Product
Action - Property
Antonym

It is possible to identify broader groupings of the above associative relationships. For example, we might distinguish broad partitive, causal, entity-property groups and a group corresponding to the AAT inter-facet relationships. However, even this simple grouping includes overlapping categories, illustrating the difficulties of creating a simple hierarchical arrangement of RT subtypes.

As an example of a highly specialised thesaurus, The UK National Railway Museum (part of the National Museum of Science and Industry) and Museum Documentation Association are constructing a railway terminology thesaurus. As part of this effort, we contributed a set of editorial guidelines on RT construction, drawing on the Getty AAT guidelines for RTs discussed earlier, which included a tailored version of the RT subtypes. The initial railway thesaurus will comprise different hierarchies but will not be a faceted thesaurus. Thus AAT inter-facet relationships were not relevant for the purposes of the thesaurus. The railway terminology editorial group agreed on a similar set of RT subtypes to the AAT guidelines, but collapsed the inter-facet relationships to one *shared operational context* and also included a *causal* relationship.

Medical information retrieval has seen a significant concentration of thesaurus related research, with influential medical thesauri like MeSH (MeSH 2000) forming part of online databases such as Medline. The US National Library of Medicine Unified Medical Language System® (UMLS 2000) is a metathesaurus bridging over 50 biomedical vocabularies including different language versions of MeSH. A metathesaurus concept has attributes, notably the higher level semantic type or category to which it belongs together with its hierarchical context in corresponding source vocabularies. Examples of semantic types are *Biologic function, Organism, Mammal.* A UMLS Semantic Network defines 54 links (relationships) between the semantic types, the most common being the *isa* relationship which establishes hierarchies. There is also a set of 5 'non-hierarchical relationships', themselves arranged in a hierarchical fashion, which may be seen as serving the function of associative relationships in the metathesaurus:

Physically related to
Spatially related to
Temporally related to
Functionally related to
Conceptually related to

There are several relationships subsumed under *Functionally related to*, of which one is *affects*, which captures various relationships associated with medical intervention or interaction of entities – causal relationships are important in the medical domain. The *affects* relationship has six children:

Manages
Treats
Disrupts
Complicates
Interacts with
Prevents

[7] distinguished from the hierarchical partitive relationship in that the terms linked belong to different categories.

Thus for the purposes of the medical metathesaurus, we have a semantic typing of concepts, more elaborate than the structure of most faceted thesauri, combined with a fairly deep hierarchy of relationship types, specialised for the purposes of the domain. Space and time are foregrounded as primary subtypes of the non-hierarchical set of relationships.

In the library domain, a subcommittee of the American Library Association (ALA) produced a Report (ALA 1999) on relationships between subjects and how they might be represented. There was a recommendation that systems should include specific relationships with a view to facilitating the development of more intelligent subject access approaches to controlled vocabulary retrieval systems. In particular, Appendix B presents a hierarchical taxonomy of relationship types, based on Michel's extensive review of the literature, with definitions of over 100 associative relationships. In the taxonomy, there are 9 first level subtypes of associative relationships:

Combined ideas
Conceptually related terms
Contiguity
Definitional associative relationships
Different hierarchy associative relationships
Meaning overlap associative relationships
Same hierarchy associative relationships
Scope issues
Unspecified associative relationships

These are in turn broken down further, to varying depth. Two major subtypes echo the distinction made in ISO2788: *Different hierarchy associative relationships* and *Same hierarchy associative relationships*, both with relatively deep hierarchies. These are partially expanded below (note this example does not show details of sublevels for all relationships)

Same hierarchy associative relationships

- Causal – subsuming
  - Dependency
  - Generic Predecessor
  - Influencing
  - Instigator/Agent-Process
  - Process-Method
  - Material-Product
- Closely related siblings
- Considered as relationships
- Coordinate ideas
- Entities studied in mutual relationship
- Part-whole
- Persons interacting in a social context
- Property
- Reciprocals
- Similarity

Different hierarchy (or facet) associative relationships

- Environmental – subsuming
  - Abstract
  - Concrete (subsuming position in time and space)
- Etymologically related
- Process issue relationships (e.g. Producer-Product), Product-Material)
- Property issue relationships (e.g. Process-Property, attribute relationships)

This large set of RT subtypes constitutes a valuable resource. It is intended to capture the rich diversity of thesaurus practice, rather than forming any structured design proposal. However, the reason for the restriction of *Causal* and *Part-Whole* relationships to *Same hierarchy* is unclear. Some subtypes appear to represent fairly weak associative relationships (eg *Combined ideas, Conceptually related terms, Similarity)*, which might well be subsumed into a generic high-level RT relationship in any attempt at mapping to a core subset of RTs for retrieval purposes. Several subtypes capture different aspects of closely, but not completely, overlapping meaning (eg, *Meaning overlap, Scope issues,* as does the AAT *Distinguished from*) and could be grouped under that broad sub-heading for retrieval purposes. A large number of relationships reflect pairings of concepts from different facets/hierarchies, which can be seen as representing different semantic categories.

6. Is an extended core set of associative relationships possible?

This (not exhaustive) review of RTs illustrates the practical difficulties in extending the current loose definition of the associative relationship to more precise hierarchies of RT relationships. Medical thesauri stand at the more complex end of a continuum of thesaurus domains, which also includes specialised smaller thesauri which may have no facet structure and only specify the 3 basic thesaurus relationships. If we include too many subtypes, then interoperability of thesaurus mapping and retrieval software may be lost. In fact, it may not prove possible to create one single extension of thesaurus relationships, but instead there may need to be different standard extensions for different domains, e.g. medical, digital library, commerce, etc.

However, one possible approach might be to aim for a limited extension of RTs at a second level and to expect domain specialisation at lower levels. This would permit some degree of interoperability in advanced thesaurus based applications involving automated traversal of a richer set of relationships for term expansion. If the three standard thesaurus relationships formed the top level of a hierarchical structure then any such new applications would retain compatibility with the large number of existing thesauri (and indexed collections) where it is infeasible to augment the core relationships.

Many of the taxonomies of RT subtypes discussed above have their origins in editorial guidelines for creating RTs, rather than being attempts at refining the semantics of RTs for retrieval purposes (our concern in this paper). Therefore, it may prove useful to logically separate practical heuristics or methods for identifying or creating RTs, such as the occurrence of one term in another's definition or Scope note, from the semantic meaning of the relationship. Documenting practical techniques for creating RTs is important but should be a separate activity. This might also reduce the need for some RT subtypes, such as *Definitional* above[8]. As a very basic illustration, our RT guidelines (derived from the AAT editorial guidelines) for non-specialist editors constructing the railway thesaurus included the checklist in Appendix 1.

More fundamentally, we can also distinguish the meaning of a relationship from the semantic category (type) of the two concepts involved. In particular, the distinction between intra and inter facet/hierarchy relationships is common but can lead to apparent illogical contrasts between and within systems. For example, should *Part-whole and Causal*-type relationships be assigned to one or the other, or both? Furthermore, in many systems, several subtypes of RT address various forms of relationships between categories represented by different facets, for example *Agent-Process, Process-Product, Material-Product,* etc.) Rather than making an a priori distinction between intra and inter facet/hierarchy relationships, it may be more useful to foreground the category of the concept (term) as an explicit aspect of the relationship in its own right. This would entail a hierarchy of semantic categories, with generic *Concepts* at the bottom level (default for terms in simple systems), belonging to various *Hierarchies* (and sub-hierarchies/minor facets) at the next level and with a set of *Facets* as the broadest level – see examples below. The semantic category of a term would be valuable to many applications automatically processing thesaurus-based metadata.

Separating out the semantic category of concepts from thesaurus relationships is useful in its own right, but could potentially yield an additional benefit. If (following the faceted approach taken by the AAT editors) the semantic category of a concept is taken as a dimension separate from the type of relationship then a smaller number of inter-facet RT relationships might suffice. The same *Causal, Uses/Requires, Spatial* or *Temporal* relationship might, at a

[8] It may be that editorial RT subtype definitions would be retained separately in editorial guidelines for constructing thesauri.

high level, connect various categories of concepts[9]. Conceivably, this might permit a restricted second level core set of RT subtypes to be applicable across some range of thesaurus domains (although this would need investigation). These relationships could themselves be refined into richer subtypes when the purposes of the thesaurus warranted. While this could be seen as shifting effort onto the identification of standard categories or facets, it can be argued that there is already a fair degree of agreement in this area.

For example, the ISO Standard refers to implicit categories of concepts which can assist an editor, say in compiling hierarchies. Examples of general categories given by the Standard include Concrete entities (such as Things and Materials), Abstract Entities (such as Events and Units of measurement) and Individual entities (proper nouns). When such categories are represented in thesauri, they are usually identified as *facets*, each facet with its own hierarchical sub-divisions. Facet analysis[10] has been a longstanding technique in thesaurus construction; concepts are decomposed into elemental classes, or facets, which form homogenous mutually exclusive groups (Aitchison and Gilchrist 1987). Faceted thesauri or classification systems include MESH, BLISS, PRECIS and the AAT. For example, the AAT (Soergel 1995) is organised into 7 facets (and 33 hierarchies as subdivisions): Associated concepts, Physical attributes, Styles and periods, Agents, Activities, Materials, Objects and optional facets for time and place. Categories such as Agent, Event, Material, Object, Time and Place are likely to be common to many thesauri. As one possible example of an extended set of associative relationships, Table 9 contains a broad grouping of RT subtypes (other groupings are also possible), which could be combined with a specification of a term's semantic category.

RT* (*Plain) for undifferentiated associative relationships

Meaning connection
Meaning overlap
Distinguished from
Antonym
Conjuncted terms

Causal (taken broadly – many inter-facet relationships might fit here)
Dependency/requires
Uses
Product
Patient
and possibly Spatial and Temporal connections

Partitive (taken broadly)
Constituent parts
Aggregate group
Property/attribute

Table 9. Example of broad groupings of RT subtypes

[9] Simplification via the intersection of different ordering principles has parallels in other domains. For example, anthropologists have investigated how cultures organise and group the cognitive principles underlying social behaviour. Tyler's (1969) classification of the underlying semantic structures observed in cognitive anthropology, included the familiar taxonomic hierarchical relationship. He also identified a 'paradigm' ordering principle, a non-hierarchical ordering which cuts across levels of taxonomic hierarchies by multiple intersections. For example, the attributes gender (male, female) and maturity (child, adolescent, adult) intersect with a mammal hierarchy to yield concepts such as mare, colt, boar, etc.
[10] The faceted approach to subject analysis began in 1933 with Ranganathan's Colon Classification (Personality, Matter, Energy, Space and Time) and was subsequently elaborated by the British Classification Research Group.

To take examples from the AAT, RT relationships between *Agents* and *Materials*, *Agents* and *Products*, *Materials* and *Objects* might all be represented by Causal subtypes in the above grouping. The semantic categories of the concepts involved further define the nature of the relationship. For example, an RT of type Causal/Uses (if the grouping in Table 9 were used) could be applied to various AAT RT subtypes, provided that additional context was provided by the semantic category of the concepts involved. An RT (of type 3Q: Activity – Equipment needed or produced) exists in the AAT between *arrows* and *archery*. A specification of the relationship would include the categories of the two concepts *(Objects/Weapons&Ammunition/arrows* and *Activities/Physical Activities/archery)*. An automated traversal application would at the least be able to ascertain that the relationship asserted that a particular kind of object was used in an activity. The relationship could be refined hierarchically if appropriate for a particular thesaurus. Similarly, a Causal/Uses RT could be employed for an AAT RT (of type 3T: Locational setting – Equipment used or produced) connecting *airports (Objects/Built Complexes & Districts)* with *aircraft (Objects/Transportation Vehicles).*

7. Conclusions

It may be impractical to expect non-specialist users to manually browse very large thesauri (for example, there are 1792 terms in the AAT's *Tools&Equipment* hierarchy). Semantic distance measures operating over thesaurus relationships can underpin interactive and automatic query expansion techniques. Ranked lists of candidate terms can assist query expansion or automatic ranking of information items in retrieval, thesaurus mapping and semantic web applications.

Online gazetteers and geographical thesauri may not contain co-ordinate data for all places and regions or, if they do, associate place names with a limited spatial footprint (eg centroid or minimum bounding rectangle). In such situations, the ability to rank places within a vicinity according to hierarchical (or other) relationships in a spatial terminology system can be useful. Section 2 provides examples of the operation of semantic distance measures over hierarchical spatial-partitive thesaurus relationships. In contexts where administrative boundaries are highly relevant, distance measures could combine quantitative and qualitative spatial relationships.

Related work has highlighted the contribution of RTs to thesaurus search aids but has noted the potential for an uncontrolled increase in result sets and a loss in precision (in cases where there is a specific search goal). Experimental scenarios (Section 4) exploring different factors relating to incorporation of RTs in semantic distance measures suggest a potential for filtering on the hierarchical context of an RT link in faceted thesauri and for filtering on subtypes of RT relationships. Specialising RTs allows the possibility of dynamically linking RT type to query context. In practice, it is likely that a combination of heuristics will be useful. In general, more control can be transferred to the retrieval system to selectively traverse RT relationship or to weight them differently. The ability in retrieval to either specialise RTs or to treat them as generic retains the advantages of the standard minimal set of thesaurus relationships for interoperability purposes, while allowing an option of a richer set of RT sub-relationships.

We have suggested the possibility of enriching the specification and semantics of RT relationships, while maintaining compatibility with traditional thesauri, via a limited hierarchical extension of the associative (and hierarchical) relationships. This would be facilitated by distinguishing the type of term from the (sub)type of relationship and explicitly specifying semantic categories for terms following a faceted approach. It may also be useful to make a distinction between heuristics for identifying/creating RTs in thesaurus construction, such as the occurrence of one term in another's definition or Scope note, from the semantic meaning of the relationship for retrieval purposes.

There are implications for thesaurus developers and implementers. A systematic approach to RT application in thesaurus design, as in the AAT, has potential for retrieval systems. Information (such as relationship subtypes) used in thesaurus design should be retained in data models and database design for later use in retrieval algorithms. Various possibilities exist for the user to characterise information need. In future work, we intend to explore utility and usability issues concerned with the incorporation of semantic distance controls in the search system user interface.

Acknowledgements

An early version of this paper was presented at the European Conference on Digital Libraries, Lisbon, 2000. We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant GR/M66233/01). We would like to thank the J. Paul Getty Trust for provision of their vocabularies and in particular Patricia Harpring and Alison Chipman for information on the AAT and Related Terms; Diana Murray and the Royal Commission on the Ancient and Historical Monuments of Scotland for provision of their dataset; Martin Doerr and Christos Georgis from the FORTH Institute of Computer Science for assistance with the SIS; and helpful suggestions from Ceri Binding and participants at the NKOS Workshop at ECDL2000. Use of the Getty Vocabularies is subject to the terms of their licenses.

References

AAT 1995. The AAT Editorial Manual: Related terms. *User Friendly*, 2(3-4), 6-15. J. Paul Getty Trust.

AAT 2000. http://www.getty.edu/gri/vocabularies/index.htm

Aitchison J., Gilchrist A. 1987. *Thesaurus construction: a practical manual.* ASLIB: London.

ALA 1999. Final Report to the ALCTS/CCS Subject Analysis Committee. Greenberg J., Hemmasi H., Kuhr P., Michel D., Riel S., Strawn G., Wool G., El-Hoshy L. http://www.ala.org/alcts/organization/ccs/sac/rpt97rev.html

Alani H., Jones C., Tudhope D. 2000. Associative and Spatial Relationships in Thesaurus-based Retrieval. *Proceedings (ECDL2000) Fourth European Conference on Research and Advanced Technology for Digital Libraries* (J. Borbinha, T. Baker eds.). Lecture Notes in Computer Science, Berlin: Springer, 45-58.

Alani H., Jones C., Tudhope D. 2001. Voronoi-based region approximation for geographical information retrieval with online gazetteers. *International Journal of Geographical Information Science.* In press.

Amann B., Fundulaki I. 1999. Integrating ontologies and thesauri to build RDF schemas. *Proc. 3rd European Conference on Digital Libraries (ECDL'99),* (S. Abiteboul and A. Vercoustre eds.) Lecture Notes in Computer Science 1696, Springer-Verlag: Berlin, 234-253.

Beaulieu M. 1997. Experiments on interfaces to support query expansion. *Journal of Documentation,* 53(1), 8-19.

Bechofer S. 2000. OIL: The Ontology Inference Layer. Special Workshop on Networked Knowledge Organization Systems. Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL2000).

Bechhofer S., Goble C. 1999. Classification Based Navigation and Retrieval for Picture Archives. *Proc. IFIP WG2.6 Conference on Data Semantics,* Rotorua, New Zealand.

Brachman R. 1983. What IS-A is and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer,* 16(10), 30-36.

Brooks T. 1997. The relevance aura of bibliographic records. *Information Processing and Management,* 33(1), 69-80.

Bruza P. 1990. Hyperindices: A novel aid for searching in hypermedia. *Proc. ACM European Conference on Hypermedia Technology*, 109-122.

Chen H., Ng T., Martinez J., Schatz B. 1997. 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,* 48(1), 17-31.

Constantopolous P., Doerr M. 1993. The Semantic Index System - A brief presentation. Institute of Computer Science Technical Report. FORTH-Hellas, GR-71110 Heraklion, Crete.

Cross P., Brickley D., Koch. T. 2000. Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema. http://www.desire.org/results/discovery/rdfthesschema.html

Cunliffe D., Taylor C., Tudhope D. 1997. Query-based navigation in semantically indexed hypermedia. *Proc. 8th ACM Conference on Hypertext,* 87-95.

Doerr M. 2000. Semantic problems of thesaurus mapping. Special Workshop on Networked Knowledge Organization Systems. Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL2000).

Doerr M., Fundulaki I. 1998. SIS-TMS: A thesaurus management system for distributed digital collections. *Proc. 2nd European Conference on Digital Libraries (ECDL'98),* (C. Nikolaou and C. Stephanidis eds.) Lecture Notes in Computer Science 1513, Springer-Verlag: Berlin, 215-234.

Fidel R. 1991. Searchers' selection of search keys (I-III), *Journal of American Society for Information Science,* 42(7), 490-527.

Guarino N. 1995. Ontologies and knowledge bases: towards a terminological clarification. In: *Towards very large knowledge bases: knowledge building and knowledge sharing,* 25-32. IOS Press.

Harper Collins, 2000, Bartholomew. http://www.bartholomewmaps.com

Harpring P. 1997. The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. *Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97),* 237-251, Archives and Museum Informatics.

Harpring P. 1999. How forcible are the right words: overview of applications and interfaces incorporating the Getty vocabularies. *Proc. Museums and the Web 1999.* Archives and Museum Informatics. http://www.archimuse.com/mw99/papers/harpring/harpring.html

Hill L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. *Proc. 4th European Conference on Research and Advanced Technology for Digital Libraries* (J. Borbinha, T. Baker eds.). Lecture Notes in Computer Science, Berlin: Springer, Hodge G. 2000. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. The Digital Library Federation Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub91abst.html

Koch T. 2000. Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review,* 24(1), 24-34.

ISO 1986. Guidelines for the establishment and development of monolingual thesauri. *ISO 2788* (BS 5723).

Jones C. 1997. Geographic Interfaces to Museum Collections. *Proc. 4th International Conference on Hypermedia and Interactivity in Museums (ICHIM'97),* 226-236, Archives and Museum Informatics.

Jones, S. 1993. A Thesaurus Data Model for an Intelligent Retrieval System. *Journal of Information Science* 19: 167-178.

Jones S., Gatford M., Robertson S., Hancock-Beaulieu M., Secker J., Walker S. 1995. Interactive Thesaurus Navigation: Intelligence Rules OK?, *Journal of the American Society for Information Science*, 46(1), 52-59.

Kristensen J. 1993. Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, 29(6), 733-744.

MeSH 2000. Medical Subject Headings. http://www.nlm.nih.gov/mesh/meshhome.html

Michard A., Pham-Dac G. 1998. Description of Collections and Encyclopaedias on the Web using XML. *Archives and Museum Informatics*, 12(1), 39-79.

Milstead J. 1999. Report on NISO Workshop on Electronic Thesauri: Planning for a Standard. http://www.niso.org/thes99rprt.html

Molholt P. 1996. Standardization of inter-concept links and their usage. *Proc. 4th International ISKO Conference, Advances in Knowledge Organisation (5),* 65-71.

Murray D. 1997. GIS in RCAHMS. *MDA Information,* 2(3): 35-38.

Pollitt A. 1997. Interactive information retrieval based on facetted classification using views. *Proc. 6th International Study Conference on Classification,* London.

Rada R., Mili H., Bicknell E., Blettner M. (1989). Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics,* 19(1), 17-30.

Rada R, Barlow J., Potharst J., Zanstra P., Bijstra D. 1991. Document ranking using an enriched thesaurus. *Journal of Documentation,* 47(3), 240-253.

Soergel. D 1995. The Art and Architecture Thesaurus (AAT): a critical appraisal. *Visual Resources,* 10(4), 369-400.

Storey V. 1993. Understanding semantic relationships. *VLDB Journal,* 2, 455-488.

Tudhope D., Taylor C. 1997. Navigation via Similarity: automatic linking based on semantic closeness. *Information Processing and Management,* 33(2), 233-242.

Tudhope D. Cunliffe D. 1999. Semantic index hypermedia: linking information disciplines. *ACM Computing Surveys,* Electronic Symposium on Hypertext and Hypermedia, 31(4es).

Tyler S. (ed.) 1969. *Cognitive anthropology.* Holt, Rinehart and Winston: New York.

UMLS 2000. Unified Medical Language System. http://www.nlm.nih.gov/research/umls/umlsmain.html

Appendix 1. Example of checklist for constructing RTs (our extension of Getty RT guidelines)

For each hierarchy, starting at top in descending hierarchical order, take each term in turn and:

1. Evaluate Scope Note for the meaning of the concept and key words representing it.

2. Consider what possible Related Terms might exist:
a) by being directly triggered from the scope note (or concept's meaning) – consider what RT subtype it would correspond to?
b) by taking each RT subtype in turn (see below) and considering if the subtype is applicable. Make a note of which RT subtype code on Spreadsheet when entering RT details.
c) (as a final step) by quickly scanning through any likely hierarchies to see if a term suggests itself

3. It is possible this may result in a strongly suggested RT term not yet in the thesaurus. If step 2 has resulted in a new candidate term for the thesaurus, write candidate term memo.

4. Check that the new RT link is not really an Equivalence or Hierarchical Relationship.

5. Check that the Scope Note of the Related Term is consistent with the relationship.

6. Check that all NT terms of the new related term are also valid for the RT relationship. The AAT followed a principle of 'inheritance' when making an RT. The RT should be made to the broadest possible related term and the relationship should hold for narrower terms of the one related (but there is no need to make an RT link to each narrower term).

7. Conversely, consider whether the source term is the appropriate term in its immediate hierarchy to make the link from. Starting from the top of each hierarchy will make the link from the broadest possible term. AAT guidelines suggest that RTs work better at 'fairly broad levels' rather than narrow level terms (exceptions including when a deep level term has no or few siblings).