

Retrieval Requirements of Faceted Thesauri in Interactive Information Systems

David Bearman

Archives & Museums Informatics, Pittsburgh, PA 15217, USA

Toni Petersen

Art & Architecture Thesaurus, The Getty Art History Information Program
Williamstown, MA 01267, USA

ABSTRACT

The *Art and Architecture Thesaurus (AAT)* will be used as the example of a faceted thesaurus that raises problems for retrieval in existing data entry and retrieval systems. This paper first describes *AAT*'s classified, faceted structure and the *AAT Application Protocol*, which is a guideline for building semantic groupings of terms into complex expressions and strings at the time of indexing. Retrieval is hampered by retrieval software in current systems that neither exploit the relationships expressed in the thesaurus nor develop the necessary syntax to search precoordinate expressions and strings. This paper identifies a number of requirements that must be met by automated retrieval systems implemented to search the *AAT* or other faceted, hierarchical thesauri. The body of the paper confines itself strictly to requirements for execution of retrieval logic and only in the final section suggests some user interface features that would be desirable for such systems.

THE ART AND ARCHITECTURE THESAURUS

The purpose of indexing is to assign terms to describe items in a collection that are most likely to match terms used in a query, and thereby retrieve records that are pertinent to the searcher. A thesaurus is designed to assist in this process, first by suggesting appropriate terms to both indexers and searchers by defining the relationships among terms and then by assisting searchers in identifying items indexed by terms different from the ones they know. Often, appropriate terms are more specific than those terms first considered. Indexers also traditionally assign the most specific applicable term. The thesaurus should help searchers locate terms that are narrower than the one they started with and sometimes, when it is necessary, to expand a concept in a search by locating broader terms. Thesauri should also help searchers by defining synonymy and associative (i.e., related term) relations and including pointers between terms bearing these relationships to one another. A system that does not help locate synonyms and narrower or broader terms would prevent rather than facilitate retrieval of records.

This position paper sets out to identify the functional requirements for a retrieval system that allows for full use of faceted thesauri to construct complex expressions and strings. A more detailed treatment will appear in a paper we intend to submit to the journal *Art Documentation*, published by the Art Libraries Society of North America.

THE ART AND ARCHITECTURE THESAURUS

The *AAT*, an operating unit within the Art History Information Program of the J. Paul Getty Trust, was developed to address the need for a standardized vocabulary of art and architecture terms for use in bibliographic and visual databases and in the documentation of object collections. The *AAT* has built its vocabulary from existing lists of subject headings augmented by the literature of the field, and as far as possible, it reflects common usage by scholars and practitioners. *AAT* terminology is prepared by an editorial staff according to American and international standards for thesaurus construction, and uses the structure of the National Library of Medicine's *Medical Subject Headings (MeSH)* thesaurus as a model. It is reviewed and approved by advisory boards composed of experts in the fields of architecture, the decorative and fine arts, archives, and information managers in these fields. The first edition of the *AAT*, comprising twenty-three of a projected forty-one hierarchies, was published in a three-volume set and on diskettes in the spring of 1990 by Oxford University Press. The *AAT* currently includes over fifteen thousand preferred terms (terms recommended for indexing) and twenty-five thousand lead-in terms (i.e., anticipated search terms composed of synonyms and variant forms of the preferred terms).

The *AAT* presents its forty-one hierarchies in a classification scheme arranged under seven facets. Facets are mutually exclusive classes of terms whose members share characteristics that distinguish them from members of other classes. The *AAT* arrived at the decision to build a faceted vocabulary due to the nature of the terminology in its domain. Since the majority of terms in the *AAT* are object names, and since it is apparent that descriptions of objects of art and architecture often contain multiword phrases that combine nouns and adjectives (designating material, style, technique, and function, among others), and that these phrases occur in infinite combinations, it was decided to provide the building blocks of these descriptions in the form of facets. These facets are: Associated Concepts, Physical Attributes, Styles and Periods, Agents, Activities, Materials, and Objects (see Figure 1).

Hierarchies may be thought of as subfacets of these original facets in that they are also homogeneous groupings of terms. In the *AAT* they are tree structures that graphically display genus-species or class-subclass relationships for a specific family of terms (see Figure 2). The hierarchical display allows for browsing broader and narrower terms and for selecting terms at appropriate levels of specificity. Because of its strict faceted structure, the *AAT* limits the hierarchical relationship to that of genus-species and avoids whole-part and polyhierarchical relationships as much as possible.

In the hierarchical display, a broader term gives the immediate class, or genus, to which the term (called a *descriptor*)* belongs and thus serves to clarify its meaning. If a term is always a type of, kind of, example of, or manifestation of another term, then it exists in the hierarchy as a narrower term to the broader term to which it is so related. Descriptors that share the same broader term are called *siblings*. Siblings are usually arranged alphabetically within their family cluster, although they occasionally may be found arranged in chronological order or by size, if such an arrangement is more appropriate. For example, chairs are always types of seating furniture but not all seating furniture is chairs. Seating furniture also includes stools and

* This paper uses *term* (unless discussed in a special context) and *descriptor* interchangeably.

benches. Therefore **seating furniture** is a broader term to the sibling terms **benches**, **chairs**, and **stools**. Similarly, in Figure 2, **dropleaf tables** are kinds of **tables by form**, but the latter includes a number of others: **capstan tables**, **cricket tables**, **draw tables**, **dropleaf tables**, and so forth. Ten siblings as directly narrower terms under **tables by form** are shown. By the way, terms in pointy brackets, such as <**tables by form**>, are used as *guide terms* and are intended to be unavailable as indexing terms; there are currently two thousand such terms.

As seen in Figure 2, hierarchies in the printed *AAT* are sorted by line number. However, the *Classification Notation (CN)* associated with each term can be found in the alphabetic index (Figure 3). For example, the **chairs** entry shows CN V.TG.AFU.AFU.ALO.AFU.AFU, the first two nodes corresponding to the notation for **Furniture** in Figure 1. The location of a descriptor in the hierarchy is determined primarily by this code, which the system generates automatically when told the parent term of the descriptor at the time it is entered into the system. The classification notation employed by the *AAT* is based on a three-letter code assigned to each level within each hierarchy. The use of three-letter aggregations provides for 26 raised to the third power, or 17,056 combinations, at any level of a hierarchy. This enables the *AAT* to distribute the vocabulary over 150 unused notations between each term at each level in the hierarchy to allow for future expansion. Classification notations enable searching of hierarchies, known as *explosion*, as discussed in the next section. The alphabetic display functions as an index to the hierarchical display by including the line number for each entry. For example, TG.20 associated with **chairs** is the sequential line number for this term in the hierarchical display.

Entries in the alphabetic index display other information about a descriptor, such as its scope note and a list of lead-in terms. The *AAT* has also instituted a new thesaurus feature called *Alternate Term (ALT)*, which is sometimes given for other grammatical forms of a descriptor when they are needed for indexing and cataloging purposes. In Figure 3 the **chairs** entry shows **ATL chair**. Alternate Terms may be used as descriptors, and the system designer must decide which form is appropriate to the application at hand and should apply the terminology consistently. For example, if a museum chooses to use the singular form of object names (**chair** instead of **chairs**), this form should be used consistently in its cataloging records.

The *AAT* vocabulary consists of descriptors that express single concepts. A single concept may be an object name (**tables** or **chairs**), a material (**white fir**), a style (**Victorian**), an activity (**cabinetmaking**), or terms from any of the other facets. These single concept descriptors are the building blocks of an indexing system. An *AAT* descriptor may be used alone or in combination to form complex modified phrases (**Victorian white fir tables**, or in syntactic combinations to form a complex string (**Victorian white fir tables—England—restoration**).

The set of procedures for coordinating *AAT* descriptors into object descriptions or indexing entries is called the *Application Protocol (Appendix D. AAT Application Protocol and Indexer's Guide* in the published *AAT*). It builds on the *AAT*'s faceted classification scheme and suggests the manner by which individual descriptors may be combined into complex expressions by adding terms in facet order:

12 PROCEEDINGS OF THE 1ST ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

- **Victorian white fir tables** = **Victorian** (facet 3) + **white fir** (facet 6) + **tables** (facet 7)
- **tinted charcoal drawings** = **tinted** (facet 5) + **charcoal** (facet 6) + **drawings** (facet 7)
- **assymetrical blue glass serving bowls** =
asymmetrical (facet 1) + **blue** (facet 2) + **glass** (facet 6) + **serving bowls** (facet 7)

There are also rules for constructing more than one expression into an indexing string. Thus, the AAT is designed specifically to support a *precoordinate indexing system*, whereby indexers apply syntactic devices to form semantic linkages at indexing time.

RETRIEVAL ISSUES

Any discussion about user expectations introduces implementation requirements for the search system. The system should take a term given by the user and look for it and many other terms (such as synonyms, narrower terms, terms from related hierarchies, or terms from across constructed headings in the database). Practical searching of this nature requires that the system employ classification notations rather than the text strings of the terms. However users must understand and be helped by the system in order to use this classification effectively. Implementation requirements are also affected by the design of the AAT to support precoordinate indexing systems (as discussed earlier). In particular, search systems should help users take advantage of the precision afforded by this indexing.

Unfortunately, automated retrieval from databases using AAT terms as envisioned by the *Application Protocol* is currently severely constrained by available information retrieval systems, especially those built for libraries and archives. Developers of online public access catalogs have provided subject retrieval functions based on the use of *Library of Congress Subject Headings (LCSH)* in MARC format. The cruder systems require the searcher to input the entire heading as given, including punctuation. Most newer systems allow some sort of keyword searching of headings, and some display an array of headings containing the search terms. In order to achieve the kind of sophisticated retrieval envisioned by the authors of the *Application Protocol*, new systems will need to be developed. This paper identifies a number of requirements that must be met by automated retrieval systems implemented to search using the AAT; these same requirements must be satisfied for searching any database indexed using faceted, hierarchical thesauri.

Priority requirements consist of the following:

1. Resolution of synonymy. A system should provide a direct substitution capability for any terms that are defined as synonyms by the thesaurus. This means that the system should link all synonyms identified with the preferred term. The satisfactory method of carrying this out is for all terms (preferred and lead-in) that represent a single concept in the thesaurus to share a common classification notation. An undesirable method for retrieving synonyms would be to search for each variant form of a term separately.

2. **Explosion of terms.** A system should provide a shorthand entry for specifying all terms in a particular hierarchy in a Boolean "or" relationship, thereby retrieving all records assigned to narrower terms than a term specified in a query. The explosion method of searching truncated classification notations increases in efficiency as the number of terms to be searched expands, because the shortest classification notations refer to the broadest terms in the hierarchy; these are generally the terms with the greatest numbers of narrower terms. This method has been used for searching hierarchies in *MeSH* for more than twenty-five years.

3. **Explosion of complex expressions.** The *AAT Application Protocol* provides indexers with instructions for constructing "expressions" that consist of *AAT* terms modified by other *AAT* terms. The resulting "expressions" are similar to precoordinated headings in *LCSH*, but without subdivisions. For example, the expression **Victorian wood tables**, although it has the appearance of a single *AAT* term, was constructed by an indexer who selected terms from different facets (Styles and Periods, Materials, and Objects, respectively), sequencing them in facet order as prescribed by the *Application Protocol*. Use of this option raises a number of additional problems that must be addressed by retrieval software. The system needs to explode each component of the expression and evaluate all term combinations according to the relationships specified in the query. Furthermore, the indexer might assign weights to components in an expression. In the current example of **Victorian wood tables**, if **Victorian** is designated the most important element with respect to the item being indexed, this item may be considered of possible interest to a searcher looking for **Victorian stained glass**. In contrast, a record for **Victorian wood tables with tables** weighted most heavily would not be of interest with respect to this search.

4. **Searching of more complex strings.** The *AAT Application Protocol* suggests that *AAT* terms can be combined in clusters of expressions to construct indexing strings, somewhat like LC subject headings containing subdivisions (e.g., **Victorian wood tables—Restoration—New York—1980**). The reasons given for constructing such strings is that they provide greater context by describing the content of an item in more detail. This prevents non-pertinent retrievals (*false drops*) that occur when several terms are assigned to the same item but are not properly associated with each other in an assigned indexing string nor presumably in the item itself. Searching these strings makes additional demands on the retrieval system. The search system will need to identify the breaks between expressions comprising the string. In the preceding example, breaks are represented by long dashes, since this string is modeled after *LCSH*. Other indexing systems might construct expressions that would link more closely the terms **Restoration**, **New York**, and **1980** since these actually form one piece of information — the restoration took place in New York in 1980. Another problem for retrieval is that the syntax of strings does not normally explicitly label the kinds of links between terms. Such links are usually inferred, much as the inference that was made concerning where and when the restoration of the Victorian wood tables took place. There is a further issue concerning strings that result from the fact that they are multiple concepts containing multiple terms. A query may pertain only to one or two of the terms within a string. All items containing strings that have those query terms should be retrieved, but the results may be quite mixed, since the strings will not match one another in all of their components. This is a complicated issue that will require user studies.

14 PROCEEDINGS OF THE 1ST ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

The foregoing requirements help define an appropriate retrieval system for an implementation that employs a faceted thesaurus in the construction of complex subject headings (expressions and strings). Two major categories of issues are raised: the way the data will have to be represented internally in the system in order to conduct the kinds of searches envisioned, and the way the user will need to perceive the system in order to be able to use it as desired.

USER INTERFACES

User interfaces for retrieval systems should support iterative construction of a query or search statement. This entails viewing not only the thesaurus itself but also being shown various results provided by the system in response to queries to the database. A multiple window environment will almost certainly be required.

The user interface should support browsing through the structure of the thesaurus while formulating a query. Users do not always have as clear a sense of the relationship between terms as is made explicit in the *AAT*. This is, of course, the primary benefit of having a formal thesaurus. But in order for that benefit to be clearly realized, a user searching, for example, **spruce furniture** needs to be able to see that **pine** is not a synonym of **spruce**, and that **pine** and **spruce** are not the only types of **softwood**. As each query term is input, the system would need to display the term and several broader and narrower terms in a separate window. When the user moves to any of these terms in the thesaurus display, a function should provide a count of the number of records in the database assigned to a term equal or below the cursor term.

Users need to be able to search single query terms, expressions, or strings in which they have been able to mark the boundaries of an expression. Recognizing that exact matches are unlikely, the system needs to report the number of usages of single terms and expressions and narrower forms of each in the database. The system will need to search each expression in the query independently against each expression in the records in the database. Users need the results ranked in order of the number of expressions for which matches were made. The system also should display results when a query only partially matches an expression. For example, a user querying **Victorian tables** would also be interested in items indexed **Victorian wood tables**, and may be interested in viewing results for **tables** as well.

CONCLUSION

The *AAT* recognizes the requirements of search systems and has developed a classification notation intended to support them. A particular concern is how to proceed in updating the thesaurus. Administrative and technical procedures are currently being defined for this, and will be issued along with the first machine-readable versions of the thesaurus. This article is intended to assist both potential users of the *AAT* who are determining how best to implement the use of this indexing language in their local systems, and the designers of information retrieval systems who are attempting to provide facilities for the searching of databases indexed with any faceted, hierarchical thesaurus.

AAT FACETS AND HIERARCHIES

(6/90)

- B **ASSOCIATED CONCEPTS FACET**
- B.BM Associated Concepts

- D **PHYSICAL ATTRIBUTES FACET**
- D.DC Design Attributes
- D.DG Design Elements
- D.DL Colors

- F **STYLES AND PERIODS FACET**
- F.FL Styles and Periods

- H **AGENTS FACET**
- H.HG People and Organizations

- K **ACTIVITIES FACET**
- K.KD Disciplines
- K.KG Functions
- K.KM Events
- K.KT Processes and Techniques

- M **MATERIALS FACET**
- M.MT Materials

- V **OBJECTS FACET**
- Built Environment*
- V.RD Settlements, Systems and Landscapes
- V.RG Built Complexes and Districts
- V.RK Single Built Works and Open Spaces
- V.RM Building Divisions and Site Elements
- V.RT Built Works Components
- Furnishings and Equipment*
- V.TB Tools and Equipment
- V.T Measuring Devices
- V.TF Hardware and Joints
- V.TG Furniture
- V.T Lighting Devices
- V.T Furnishings
- V.T Personal Artifacts
- V.T Containers
- V.T Culinary Artifacts
- V.T Musical Instruments
- V.T Recreational Artifacts
- V.T Armament
- V.T Transportation Artifacts
- V.T Communication Artifacts
- Visual and Verbal Communication*
- V.VB Image and Object Genres
- V.VD Drawings
- V.V Paintings
- V.V Prints
- V.VJ Photographs
- V.V Sculpture
- V.V Multi-Media Art Forms
- V.V Communication Design
- V.V Exchange Media
- V.V Book and Writing Forms
- V.VW Document Types

- W. Proper Names
- Y. Dates
- Z. Places

Figure 1. Sample Classified Display of AAT Facets and Hierarchies.

	<furniture by form or function> <support furniture> stands <stands by function>	FURNITURE (TG)
TG.638	drinking stands	
TG.639	globe stands	
TG.640	hall stands	
TG.641	hat stands	
TG.642	kettle stands	
TG.643	lecterns	
TG.644	map stands	
TG.645	muffin stands	
TG.646	music stands	
TG.647	plant stands	
TG.648	ferneries	
TG.649	jardinières	
TG.650	portfolio stands	
TG.651	shaving stands	
TG.652	smokers	
TG.653	smoking stands	
TG.654	torchères	
TG.655	tray stands	
TG.656	umbrella stands	
TG.657	washstands	
TG.658	basin stands	
TG.659	corner basin stands	
TG.660	corner enclosed basin stands	
TG.661	enclosed basin stands	
TG.662	<steps: furniture>	
TG.663	bed steps	
TG.664	library steps	
TG.665	tables	
TG.666	<tables by form>	
TG.667	capstan tables	
TG.668	cricket tables	
TG.669	draw tables	
TG.670	dropleaf tables	
TG.671	butterfly tables	
TG.672	fly tables	
TG.673	gate-leg tables	
TG.674	handkerchief tables	
TG.675	Pembroke tables	
TG.676	harlequin tables	
TG.677	pillar and claw Pembroke tables	
TG.678	Sutherland tables	
TG.679	end tables	
TG.680	grotto tables	
TG.681	horseshoe tables	
TG.682	kidney tables	
TG.683	occasional tables	
TG.684	tables ambulantes	
TG.685	tables de salon	
TG.686	ratonas	



May be used in combination with other descriptors (e.g., mahogany + pedestals; turned + legs; serpentine-front + chests of drawers; Federal + pier tables).

Figure 2. Sample Display of AAT Hierarchies.

Cha-Cha

chaines			
RT.630		ing the piece into a table. Chairs of this type were made from the Middle Ages to the late 17th century in Europe and until the late 19th century in America. (DDA)	chairs, Barcelona USE Barcelona chairs
ALT chaine			chairs, barrel USE barrel chairs
SN Type of wall decorations consisting of vertical bands of rusticated masonry which divide the surface into panels or bays; common in 17th-century French domestic architecture. (HAS)		UF benches, monk's (chair-tables) chairs, monk's (chair-tables) monk's benches (chair-tables) monk's chairs (chair-tables) monk's seats seats, monk's table-chairs	chairs, barrel (circular easy chairs) USE circular easy chairs
CN V.RT.AFU.AFU.BIQ.BUE.ALO. AFU.ARI.ALO		CN V.TG.AFU.AFU.ALO.AFU.AFU.AFU. DOI	chairs, Barwa USE Barwa chairs
chains, sash			chairs, basket USE basket chairs
USE sash chains			chairs, bath USE bath chairs
chair-back settees		chair vises	chairs, beach USE beach chairs
TG.301		TB.1076	chairs, beanbag USE beanbag chairs
ALT chair-back settee		ALT chair vise	chairs, bended-back USE bended-back chairs
SN Small sofas formed of two or three combined chair backs, with arms, backs, and legs similar to those of the open-back chairs of the particular period. (FAIRB)		UF vises, chair	chairs, bent-wire USE ice-cream parlor chairs
UF bar-back sofas		CN V.TB.AFU.BCW.CFS.ALO.ALO.CXA. ALO	chairs, bergere USE bergeres
chairs, double		chair web saws	chairs, Bertoua USE diamond chairs
Darby and Joan settees		USE web saws	chairs, birdcage Windsor USE square-back Windsor chairs
double chairs		chairmakers	chairs, board USE board chairs
settees, chair-back		HG.560 (A,L)	chairs, Boston USE Boston chairs
settees, Darby and Joan		ALT chairmaker	chairs, Boston rocking USE Boston rockers
sofas, bar-back		CN H.HG.AFU.AXC.BUE.AXC.BUE.ALO	chairs, Bouneparte USE Trafalgar chairs
CN V.TG.AFU.AFU.ALO.ALO.BUE.AFU		chairs	chairs, bow-back Windsor USE bow-back Windsor chairs
chair-beds		TG.20 (A,L,N,B,R)	chairs, Breuer USE Cesca chairs Wassily chairs
USE bed chairs		ALT chair	chairs, Brewster USE Brewster chairs
chair bits		SN Seats for one person with a back or a back and arms. Distinct from stools which have no back. (DDA)	chairs, Brno USE Brno chairs
USE spoon bits		CN V.TG.AFU.AFU.ALO.AFU.AFU	chairs, burgomeister USE corner chairs
chair-lift stations		chairs, abbots'	chairs, butterfly USE Hardoy chairs
USE ski-lift stations		USE Glastonbury chairs	chairs, cabinet commode USE cabinet chairs
chair lifts		chairs, Adirondack	chairs, cabriole USE cabriole chairs
RT.1297 (N)		USE Adirondack chairs	
ALT chair lift		chairs, angle	
SN Use for inclined lifts installed in buildings, usually on a staircase, and including a chair in which the passenger sits.		USE corner chairs	
UF lifts, chair		chairs, ax	
CN V.RT.ALO.ALO.AFU.AFU.ALO.AFU		USE ax chairs	
chair lifts (ski lifts)		chairs, back-stool	
USE ski lifts		USE backstools	
chair rails		chairs, balloon-back	
RT.650		USE balloon-back chairs	
ALT chair rail		chairs, balloon-back Windsor	
SN Horizontal strips, usually of wood, affixed to walls at a height which prevents the backs of chairs from damaging the wall surface.		USE balloon-back Windsor chairs	
UF rails, chair		chairs, banister	
CN V.RT.AFU.AFU.AFU.BIQ.BUE.ALO. AFU.BCW.ALO		USE banister-back chairs	
chair-tables		chairs, banister-back	
TG.47 (N)		USE banister-back chairs	
ALT chair-table		chairs, bar-back	
SN Armchairs with large backs which are hinged so that they can be swung forward to rest on the arms, convert-		USE bar-back chairs	
		chairs, barber's	
		USE barber's chairs	
		chairs, barber's (corner chairs)	
		USE corner chairs	

Figure 3. Sample Display of AAT Alphabetic Index.

