Building and Accessing Vocabulary Resources for Networked Resource Discovery and Navigation

Joseph A. Busch

GETTY VOCABULARY RESOURCES OVERVIEW

The Getty has a lengthy history in the research and development of thesauri and other structured vocabulary tools to make the use and exchange of electronic information easier. These tools include the *Art & Architecture Thesaurus* (1994) (*AAT*—a thesaurus of art-historical terminology that reflects the "common usage" of scholars and catalogers); the *Union List of Artist Names* (1994) (*ULAN*—a database of artist and architect names in both standard and variant forms along with biographical and bibliographic data); and the *Thesaurus of Geographic Names* (*TGN*—a thesaurus of worldwide geographic names and related historical and other information organized into hierarchies). Table 1 summarizes and compares the scope, coverage, and structure of the Getty vocabularies.

Table 1.
Comparison of the Scope and Coverage of Getty Vocabularies

Project	*Scope*	*Coverage*	*Sources*	*Structure*
Art & Architecture Thesaurus (AAT)	European: late antiquity—American: European discovery—Global: modern	Visual art, architecture, and material culture	Published	125,000 terms in 33 hierarchies in 7 facets
Union List of Artist Names (ULAN)	Global: antiquity—present day Western-oriented (now)	Artists, architects, engravers, etc.	Published and unpublished (archival)	200,000 clustered names
Thesaurus of Geographic Names (TGN)	Global: current and historical	Geo-political place names	Published and unpublished (archival)	300,000 names in poly-hierarchies

The *Art & Architecture Thesaurus*, begun in the early 1980s, was inspired by new subject indexing tools being designed explicitly for online resources (such as *Medical Subject Headings — MeSH*) (National Library of Medicine, 1997), but the *AAT* has been unconstrained by the operating concerns of an abstracting and indexing service. Through an advisory and review process, headings and terms related to visual art, architecture, and material culture were culled from the *Library of Congress Subject Headings* (*LCSH*) (Library of Congress, 1996) and other existing lists, reorganized into hierarchies, filled in, and extended according to the ANSI/ISO standards for thesaurus construction (National Information Standards Organization, 1994). The first edition of the *AAT* was published in 1990 with supplements, electronic versions, and tools for browsing the *AAT* following (*Art & Architecture Thesaurus*, 1994, 1990; *Art & Architecture Thesaurus: Authority Reference Tool*, 1994). The *AAT* now contains approximately 125,000 terms organized into thirty-three hierarchies in seven facets. New terms and change requests are submitted through a candidate term and comment process; scope notes and related term links continue to be added by the *AAT*, and trained experts are working on projects to develop specialized terminology areas, such as conservation, and on a variety of translation projects.

The *Union List of Artist Names* (*ULAN*) (1994) was created by clustering the artist and architect names from the authority files of nine Getty bibliographic, archival, and object record databases. The *ULAN* preferred or entry form was selected algorithmically based on the common practice among the contributor files with the *Bibliography of the History of Art* (*BHA*), the default (except for records to which the Getty Vocabulary Program made a specific research contribution). The clustering was done by merging the authority files, followed by research on each name (or name cluster) in reference sources. The first edition of the *ULAN* containing about 200,000 names was published in 1995 including an electronic edition and tools for browsing it (*Union List of Artist Names: Authority Reference Tool*, 1994). New names are submitted through a candidate process (sample MARC and tagged ASCII format data files available from: http://www.gii.getty.edu/pub/ulan/).

The *Thesaurus of Geographic Names* (*TGN*) was generated by merging authority files from Getty databases and several commercial sources (data files initially licensed from commercial sources are in the process of being replaced to ensure that the *TGN* can be freely distributed and enhanced in the future). While conventional geographical information system (GIS) databases contain geopolitical information that is based on a fixed point in time (usually the present), the *TGN* data structure holds instances of information about a location that are linked to a period of time. The *TGN*, which now contains about 300,000 geographic names, is available on the World Wide Web (http://www.gii.getty.edu/tgn_browser/).

While initially conceived as indexing tools, the Getty vocabularies were not designed specifically to support a particular operating index. They were designed to coordinate or map variant indexing practices across resources. As summarized in Table 2, the vocabularies provide a clustering of variant or synonymous term or name forms, roles associated with the "term," as well as a mapping of the term forms in pre-coordinated phrases and strings as "found" or used in a variety of sources. The vocabularies also provide hierarchic (parent-child) and associative (related term/historical) relationships which are useful in expressing a search statement. Examples of these relationships from each vocabulary are illustrated in Table 3.

A.K.A.—A WEB SEARCH REDIRECTION TOOL

The Getty has been exploring how vocabularies can be used to help generate search terms, particularly when searching across multiple databases or in unfamiliar resources. An experimental search tool—*a.k.a.*—can use the vocabularies to help generate queries against a collection of one or more of twenty-two databases and two Internet search engines—AltaVista and Lycos (a public version of *a.k.a.,* which includes four Getty databases, is accessible from the GII home page [Getty Information Institute, 1997a]). Most of the databases that include bibliographic, archival, and museum object records are mounted locally as sets of text documents

TABLE 2
COMPARISON OF THE GETTY VOCABULARY DATA TYPES

Data Types	*AAT*	*ULAN*	*TGN*
Entry term	entry term (determined by rules)	entry name (determined by contributor)	entry name (determined by contributor)
Variants	used for (UF), alternate, UK term, UK alt.	linguistic, historical, & orthographic variant(s)	linguistic, historical, & orthographic variant(s)
Role	facet, hierarchy	role, nationality, date(s), & style/genre	place type(s), date(s)
Source terms/names	source term(s), sources(s)	source(s) consulted (found/not found)	source(s)
Links	broader term (BT), related term (RT)	entity relationships (e.g., related to, student of,influenced by, etc)	current part, current whole, historical part, historical whole, related to (RT)
Notes	scope note	descriptive note	descriptive note (e.g., historical gloss)
Other	history note(s), indexing note		coordinates

TABLE 3
EXAMPLES OF THE GETTY VOCABULARY SYNDETIC RELATIONSHIPS

Relationships	*AAT*	*ULAN*	*TGN*
Synonym Clusters	bergeres UF: barjaires; barjairs; bergere chairs; bergiers; bijairs; bujairs; burjair; burjaires; burjairs; cabriole bergeres; fauteuils a panneaux; fauteuils en bergere	Giambologna (Jean Boulogne) Bologna, Giovanni; Bologna, Giovanni da; Bologna, Jean; Bologna, Jean de; Boulogne, de Jean; Giambologna; Jean Boulogne; Jean de Bologne	Wien Vienna; Vienne; Vindobna (historical); Vindobona (historical); Vindomana (historical); Bec (historical)
Roles	bergeres Furnishings Hierarchy; Objects Facet	Giambologna (Jean Boulogne) (Flemish (Italian School), 1529-1608); (Flemish sculptor in ITA, 1529-1608); (Italian artist, 1529-1608); (Italian artist, c. 1524-1608); (Italian sculptor (b. in Netherlands), 1529-1608); (Sculptor, goldsmith, 1529-1608)	Wien inhabited place, city; national capital; state capital; river port; industrial center; transportation center; cultural center; educational center; episcopal see; noble residence (historical); municipium (historical)
Source Terms	ghost towns AVERY: Cities and town—Ruined, extinct, etc. LCSH: Cities and towns, Ruined, extinct, etc. RIBA: Villages: lost	Giambologna (Jean Boulogne) Avery & Radcliffe, GIAMBOLOGNA... (1978); Dhanens, JEAN BOULOGNE	not applicable (all names in the database are fully incorporated)
Hierarchy	bergeres BT: armchairs SIB: elbow chairs...; great chairs...	not currently applicable	Wien current part: Schönbrunner Schlosspark current whole: Wien State historical part: Austro-Hungarian Empire
Associative	sinopie RT: frescoes	Giambologna (Jean Boulogne) student of Jacques Dubroeucq	Austro-Hungarian Empire ally of Germany

and searched using the WAIS (Wide Area Information Server) text search engine. Queries are made directly to several databases of full-text source materials from the University of Southern California—i.e., USC Ethnic Studies and Photo collections, one museum object record database from the Fowler Museum, and the AltaVista search engine. The vocabularies are mounted as a Sybase database. Searches can be executed directly, enhanced by searching *AAT* or *ULAN* through a series of Web pages with automatically generated results displays, or automatically expanded with *AAT* terms. When the search negotiation is completed, the final search is executed sequentially on the target databases, and the results from each one are presented as a separate set.

Some examples of the kinds of user queries that can be answered using *a.k.a.* search negotiations are: (1) use the *ULAN* to look for information about a Mexican painter named Siqueiros without knowing exactly how to spell his name; (2) use the *AAT* to look for information about freso "underdrawings" without knowing the appropriate technical term; (3) use the *ULAN* to look for information about Georgia O'Keeffe (whose name is often misspelled as O'Keefe); (4) use the *AAT* to look for information about a particular type of armchair without knowing the appropriate technical term; (5) use the *AAT* to look for the technical name of "containers" often found in Egyptian tombs; or (6) use the *ULAN* to look for information about a Greek sculptor named "Praxiteles" without knowing how to spell his name.

WEBART—A VOCABULARY RESOURCE SEARCH AND BROWSING TOOL

A problem the Getty has been facing is how to refresh and maintain the vocabulary tools. Like networked resources, these vocabulary tools are global in scope, that is to say, they are potentially infinite in terms of content depth and linguistic breadth. The Getty strategy has been to develop methodologies and tools that will enable the extension of vocabularies through network-based infrastructures as part of an overall distributed database initiative known as DDI (Distributed Database Initiative) (available on the World Wide Web at: http://www.rlg.org/strat/projahip.html).

The *a.k.a.* environment consists of two database infrastructure components—the target databases and the vocabulary databases. Uncoupling the vocabulary databases from the *a.k.a.* application provides a convenient means of providing access to *AAT*, *ULAN*, and (in the future) *TGN* and may serve as a core architecture for a future vocabulary server. This application, which is called WebART (or Web Authority Reference Tool), will replace the previous DOS terminate and stay resident application used to distribute earlier electronic versions of the AAT and ULAN. The components of the Getty's WebART tool are a Sybase database engine that holds

the vocabulary databases, a Web-interface to search it, and scripts to format the output in HTML. Hypertext links are also provided to Web pages containing instructions on use and other information about the vocabularies as well as a link to send comments to the vocabulary editor. WebART is now freely available at the Getty Information Institute (1997) Web site (http://www.rlg.org/strat/projahip.html).

The Research Libraries Group Art and Architecture Group members recently recommended that RLG develop the infrastructure for a vocabulary server that includes some of the components envisioned below. The Canadian Heritage Information Network is also involved in developing such infrastructures. The replication of vocabulary server infrastructures is desirable but also raises record flow and vocabulary authentication issues which are beyond the scope of this discussion.

FUTURE VOCABULARY CONTRIBUTION ENVIRONMENT

The vocabulary server needs to support network-based contributions of candidate terms to the various vocabularies. This would be accomplished through the implementation of a variety of registry services which may be limited to registered and authorized contributors. Contributions would automatically be posted to the vocabulary server database, graphically be identified as contributions (e.g., with a different color or font), and generate a transaction to a vocabulary maintenance system (the vocabulary link registry and management tool) for authentication by the appropriate vocabulary editor. Applicable editorial rules (for scope, literary warrant, etc.) would be accessible by hypertext links from each registry form. The following registry services, illustrated in Figure 1, are envisioned in the vocabulary server infrastructure.

—*Add associative links.* A registry mechanism would be provided to register candidate associative links (related/historical) terms and to "type" these links. The registry would be accessible from the full term record display. Source and target terms would automatically be validated.

—*Add equivalent links.* A registry mechanism would be provided to register and "type" synonym/variants (including UF, ALT, source terms, historical variants, etc.) to an existing term.

—*Add notes.* A registry mechanism would be provided to register and "type" notes (including scope and biographical notes, historical glosses, and so on) to an existing term.

—*Add links to other resources.* A registry mechanism would be provided to enter links to network sources such as Web pages, images of objects, and so on to an existing term. This could provide a framework for an ART "Yahoo" or vocabulary resource for indexing art and culture on the Web.

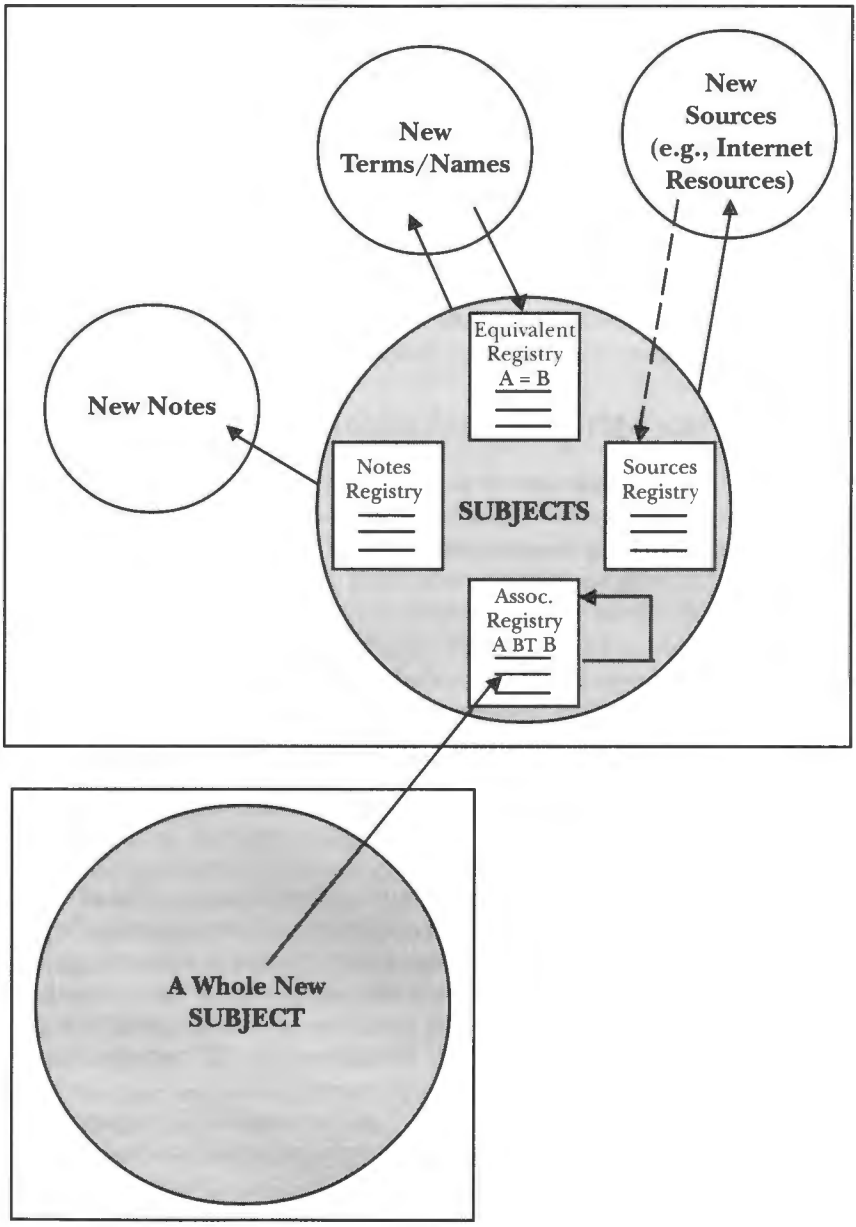

Figure 1. An Architecture for Vocabulary Contribution

—*Add a whole new subject.* The response for searches which return no hits would include the option to collect candidate term information with a Web form. The form would validate that the candidate term was unique, that the broader term was in the thesaurus, and that all required fields in the form had been completed. The contributor would be identified automatically (e.g., by their e-mail address).

ENDNOTE ON VOCABULARY SERVERS

The vocabulary server architecture is based on a cooperative model to maintain and extend a public information utility through voluntary contributions. The weakness of this model is that generosity and consensus do not necessarily lead to effective resources and efficient choices for application of effort. There are good arguments for taking a more rational approach through the application of statistical analysis techniques on the art and culture domain (see Chen et al., 1996; Schatz, 1995). For example, an analysis of the frequencies and rates of term co-occurrences in a corpus consisting of the title, abstract, and subject fields of art and architecture database records may be a promising method for identifying those terms that do not yet already exist in the vocabulary repertoire (for example, see Buckland et al., 1993, pp. 311-19).

REFERENCES

Art & architecture thesaurus (2d ed.). (1994). New York: Oxford University Press.

Art & architecture thesaurus [Electronic Edition]. (1990). New York: Oxford University Press.

Art & architecture thesaurus: Authority reference tool (version 2.1) [MS-DOS program]. (1994). New York: Oxford University Press.

Art & architecture thesaurus supplement 1. (1992). New York: Oxford University Press.

Buckland, M.; Norgard, B.; & Plaunt, C. (1993). Filing, filtering, and the first few found. *Information Technology and Libraries, 12*(3), 311-319.

Chen, H. C.; Schatz, B. R.; Ng, T.; Martinez, J.; Kirchoff, A; & Lin, C. T. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval–the Illinois Digital Library Initiative project. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 771-782.

Getty Information Institute. (1997a). *a. k. a.* Retrieved March 2, 1998 from the World Wide Web: http://www.gii.getty.edu/aka/

Getty Information Institute. (1997b). *Art & architecture thesaurus browser.* Retrieved March 2, 1998 from the World Wide Web: http://www.gii.getty.edu/aat_browser/

Getty Information Institute. (1997c). *Union list of artist names browser.* Retrieved March 2, 1998 from the World Wide Web: http://www.gii.getty.edu/ulan_browser/

Library of Congress. (1996). *Library of Congress subject headings* (19th ed.). Washington, DC: Library of Congress.

National Information Standards Organization. (1994). *Guidelines for the construction, format, and management of monolingual thesauri: An American national standard* [Z39.19-1993]. Bethesda, MD: NISO Press.

National Library of Medicine. (1997). *Medical subject headings: Annotated alphabetic list, 1997.* Bethesda, MD: National Library of Medicine.

Research Libraries Group. (1997). *Distributed database initiative–RLG and Getty Information Institute Partnership.* Retrieved March 2, 1998 from the World Wide Web: http://www.rlg.org/strat/projahip.html

Schatz, B. R. (1995). Information analysis in the net: The interspace of the twenty-first century. In *America in the age of information: A forum.* Washington, DC: Committee on Information and Communications of the National Science and Technology Council. Retrieved March 2, 1998 from the World Wide Web: http://www.hpcc.gov/cicc/cic_forum_v224/cover.html

Union list of artist names. (1994). New York: G. K. Hall.

Union list of artist names: Authority reference tool (version 1.0) [MS-DOS program]. (1994). New York: G. K. Hall.