5. Using Multiple Vocabularies

Catalogers of art information require multiple vocabularies because no single vocabulary provides the full set of terminology needed to catalog or index a given set of cultural heritage data; therefore, a combination of vocabularies is necessary for indexing. Furthermore, separate vocabularies may be required for retrieval; ideally, retrieval vocabularies are based on indexing vocabularies but may be optimized and applied differently for this purpose. Strategies for using vocabularies for indexing and for retrieval are further discussed in **Chapter 8: Indexing with Controlled Vocabularies** and **Chapter 9: Retrieval Using Controlled Vocabularies.**

In order to overcome the obstacles involved with using multiple vocabularies, systems developers should investigate the interoperability of vocabularies and the creation of local authorities.

5.1. Interoperability of Vocabularies

In the context of controlled vocabularies, *interoperability* refers to the ability of two or more vocabularies and their systems or components of their systems to map to each other's data, with the goals of exchanging information and enhancing discovery. Interoperability of controlled vocabularies is a complex topic that has been researched in the field of information science since the 1960s.

Interoperability deals with the two conflicting demands that underlie the development and use of controlled vocabularies. The first demand is that specialized vocabularies be developed for a certain community, such as the art and cultural heritage community; these vocabularies reflect the specific terms and concepts needed by catalogers to index and classify that material. However, no single vocabulary can be comprehensive, not even for its given scope. Interoperability may thus come into play as catalogers assign indexing terms to material, because cataloging art information requires a broad range of terminology that comes from different sources.

The second demand is made by end users who want to use a single search to find resources (e.g., texts, data, images, etc.) in federated

settings across resources in different domains and created by different communities. Interoperability between resources and vocabularies is also a critical factor in meeting this demand.

Mappings between vocabularies may be used to facilitate faster indexing when two or more vocabularies are used by the indexer. When the indexer selects a term from the first vocabulary, the system can respond by offering corresponding terms from the second vocabulary. The indexer then confirms appropriate selections and rejects those that do not apply. In addition, creating interoperability between vocabularies for retrieval can expand retrieval options for a given collection without the cost of additional indexing by indexers having to select terms from the second vocabulary.

5.2. Maintenance of Mappings

The use of multiple controlled vocabularies across multiple databases and systems involves the mapping of terms and the design of methods to use those terms for indexing and retrieval. In addition, it requires plans for maintenance of the vocabularies and the mapping; terminologies tend to change significantly over time, thus rendering the mapping obsolete if a maintenance plan is not in place.

The issues surrounding interoperability are discussed in detail in *ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies; BS 8723-4:2007: Structured Vocabularies for Information Retrieval: Interoperability between Vocabularies;* and *ISO/CD 25964-1: Thesauri and Interoperability with Other Vocabularies. Part 1: Thesauri for Information Retrieval* (in development at the time of this writing). A brief discussion of the issues appears below. Additional issues surrounding retrieval using vocabularies are addressed in **Chapter 9: Retrieval Using Controlled Vocabularies**.

5.3. Methods of Achieving Interoperability

Achieving interoperability requires adapting two or more vocabularies—which were probably developed to stand alone—to work in a new environment where search terms drawn from one link to terms found in the other. Often the search is conducted across two or more resources. The resources may have been indexed using one, all, or none of the vocabularies being used in retrieval.

Thus, interoperability may involve merging or adapting two or more controlled vocabularies to actually or virtually form a new controlled vocabulary that combines all the concepts and terms contained in the originals. It could also involve merging or adapting

two or more resources that have been indexed using different controlled vocabularies. Various methodologies for direct mapping and switching may be used.

5.3.1. Direct Mapping

Direct mapping generally refers to the matching of terms one-to-one in each controlled vocabulary. The vocabularies need not be the same size (one may be smaller or larger) or cover exactly the same content, but there should be significant overlap in content. This technique assumes that where overlap exists, there is the same meaning and level of specificity between the two terms in each controlled vocabulary. In the broadest application, interoperability allows vocabularies developed for completely different domains to be combined in a comprehensive conceptual and terminological map. Successful mappings typically begin with a master vocabulary to which one or more subsidiary vocabularies are mapped, rather than mapping back and forth across both or all vocabularies.

Mapping may be done by computer algorithm or human mediation, but often both methods are employed together. The advantage of human mediation in creating mappings is that a subject expert can make a judgment about inexact equivalents. However, the use of automation or partial automation in a first pass at mapping may be beneficial.

Automated mapping may employ sets of terms found through comparisons and analysis. In one example, *co-occurrence mapping,* a set of terms may be created based on clusters of related terms gathered from the target resources. Related terms are determined by the frequency with which the terms appear together in the data. The result is a body of sets of presumably loosely related terms. The terms used for the co-occurrence mapping may be selected from individual metadata fields in the resources, from uncontrolled keywords assigned to the content or from the full text of the content in the resources. The loosely mapped term clusters discovered via this approach may be used in mapping between controlled vocabularies or used directly for indexing and retrieval.

In another automated strategy, links between vocabularies may be made through a temporary union list created dynamically in response to user queries. Such algorithms may map terms that are not necessarily conceptual equivalents but may be related in some way and may be used to map to existing controlled vocabularies. Capturing these clusters of presumably related terms is intended to enhance indexing and retrieval at the time a user enters a query, but no new controlled vocabulary is permanently generated.

5.3.2. Switching Vocabulary

Switching refers to the use of a third vocabulary, a switching vocabulary, that itself can link to terms in each of the two original controlled vocabularies. As with direct mapping, this type of mapping also assumes that the meaning of the terms can be reconciled—in this case, between all three terms: the original two controlled vocabulary terms and one switching term. The advantage of this method is that the scope and format of the switching term may be made broad enough to compensate for differences between the two original terms. Another application of switching occurs when the third vocabulary provides notations or a classification scheme under which terms from both original controlled vocabularies may be grouped. For example, *carriage cradles* in one vocabulary and *swinging cradles* in a second vocabulary could both be mapped as children of *cradles* in a switching vocabulary. This approach enables a single, unifying hierarchical display for terms that originated in multiple sources.

A further example of using a third vocabulary to map two or more original vocabularies involves a *lexical database.* This kind of database can be used to link terms from multiple controlled vocabularies into clusters of related concepts for which the types of relationships are defined, such as synonyms, antonyms, hierarchical relationships, and associative relationships.

5.3.3. Factors for Successful Interoperability of Vocabularies

The achievement of interoperability depends upon various factors, including the following:

Scope of mapping: The greater the number of elements included in the mapping, the more difficult the mapping becomes. At minimum, a mapping between vocabularies should match terms to terms. If a mapping intends to link not only terms but also scope notes, relationships, and other elements of the records from each vocabulary, more human intervention is required to harmonize the results.

Similarity of content: The more similarity there is in the content of each of the vocabularies and of the resources being searched, the more likely it is that successful interoperability will be achieved. For example, since there is little overlap in the content, trying to map an art vocabulary to a medical vocabulary for indexing and retrieval purposes has little advantage over using each vocabulary separately in indexing and retrieval. Even when both controlled vocabularies comply with standards

such as those from ISO or NISO thesaurus standards, if the content is not similar, differences and variability in terminology, meaning, and syntax will hamper cross-domain interoperability.

Intended audience: If the purposes or intended audiences of the resources or vocabularies are very different, mappings of vocabularies are difficult or impossible and search results are uneven. If one database is indexed using terms for nonspecialists while the other is indexed for subject experts, users from both communities are likely to be disappointed with the combined retrieval results. For example, the resources and vocabularies required for an audience of K–12 students typically differ from those required for scholars and subject experts.

Format and hierarchical structure: The more there is similarity in the format and hierarchical structure of the vocabularies, the more likely interoperability between them is successful. If terms from the different vocabularies vary in format and hierarchical structures, indexing and retrieval results may be poor, even when the combined vocabularies are similar in content and used to search across similar domains. For example, mapping subject headings to thesaurus terms is typically only marginally successful, because subject headings are made of multiple terms and other information—such as dates—concatenated together, usually without hierarchical structure, while each term in a thesaurus is a single word or short phrase representing a discrete concept that is organized in a strictly defined hierarchical context. Interoperability between two or more such controlled vocabularies usually must reduce or eliminate structure while attempting to maintain meaning, which is difficult with a thesaurus because meaning is implied by the hierarchical context of the term.

Precoordination and postcoordination: Differences in the application of precoordinated and postcoordinated terminology in the vocabularies complicate mapping efforts if one vocabulary contains headings while the other contains unique terms. For example, a two-to-one match rather than a one-to-one match is required for the heading *Baroque cathedral* if the second vocabulary places the style *Baroque* in one hierarchy and the building type by function, *cathedral,* in a second hierarchy.

A related issue concerns the differences in precoordination and postcoordination expected in the search

methodologies of the resources being searched; if one database is indexed for precoordinated terms and the second expects terms to be postcoordinated in retrieval, results are uneven. Libraries have agreed on a common search protocol—*Information Retrieval: Application Service Definition and Protocol Specification* (*ANSI/NISO Z39.50*)—to perform searches across multiple Online Public Access Catalogs (OPACs). More recently developed search protocols are *Search/Retrieve via URL* (*SRU*), *Search Retrieve Web Service* (*SRW*), and *Metasearch XML Gateway* (*MXG*). However, resources in other communities do not typically have a common protocol, causing challenges in the interpretation of search terms and search results.

Granularity and specificity: The differences in degree of specificity or granularity of the controlled vocabularies themselves, and of the indexers' applications of the vocabularies in the target resources, may result in uneven results in indexing and retrieval. For example, if one vocabulary contains very specific terms for a given domain while another contains only general terms, mapping between them will be difficult. If an exact equivalent is not available, mappings should attempt to link to broader terms, narrower terms, or to terms that have overlapping, if not synonymous, meaning.

Conversely, if indexers of both resources have used the same vocabulary for indexing, even if they use varying degrees of specificity and granularity in indexing terms, retrieval using that vocabulary across resources is still likely to be relatively successful because the broader and narrower terms are logically linked in the vocabulary and may be applied together in a search.

Synonymy and near synonymy: Differences in how synonyms and near synonyms are handled affects the ability to make a successful mapping between vocabularies. If one vocabulary links near synonyms as *used for* terms for a concept, while the other links only true synonyms, it is difficult to make a one-to-one match between concepts. For example, *levitation* and *flight* may be related in a very general way and could both be terms in a single thesaurus record, but they are not true synonyms because their meanings are different, thus they comprise two separate records in a thesaurus employing only true synonymy.

Authoritativeness: If vocabularies differ in the level of authoritativeness by which they are developed, mapping them is difficult. For example, if the literary, organizational, and user warrants allowed in developing the various vocabularies are quite different, there may be little commonality among the terms across the vocabularies or different meanings for the same term.

5.3.4. Semantic Mapping

A *semantic network* comprises relationships between terms and concepts based on their meanings or the nature of the relationships between them. The semantic relationships are sometimes derived from the vocabularies. In other cases, they are extrapolated from the target content databases.

A semantic network may be used to map terms from one or more controlled vocabularies according to a defined underlying organizational structure or conceptual scheme. The relationships may range from a simple hierarchical structure with generic broader/narrower relationships to a more complex set of carefully defined relationships, such as *contained in, agent for, process is,* etc. The relationships may be categorized to indicate the degree of closeness between linked terms, for example

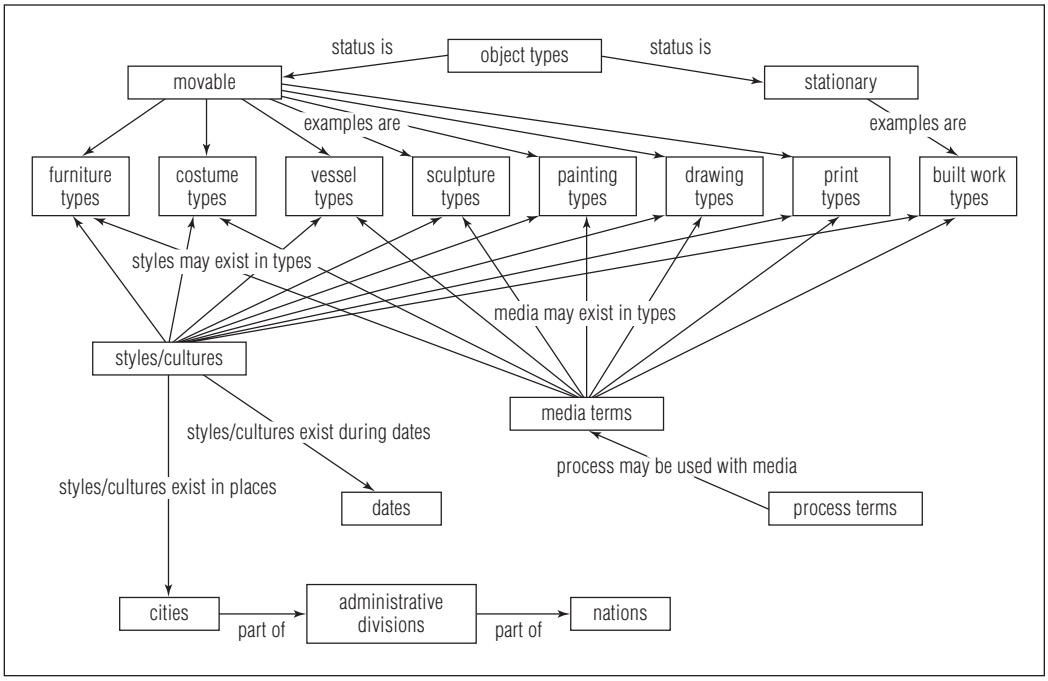

Fig. 35. Diagram of an example generic semantic mapping for artworks, where elements in rectangular boxes are linked to other elements using the relationships designated (e.g., *status is*).

exact synonyms, near synonyms, closely related terms, loosely related terms, and *antonyms.*

A semantic mapping based on categories and relationships is illustrated in the diagram on the previous page. See also the discussion of ontologies in **Chapter 2: What Are Controlled Vocabularies?**

5.4. Interoperability across Languages

Multilingual controlled vocabularies are sometimes treated as a special case of interoperability. If unique vocabularies have been developed independently using different languages, utilizing the two together as a multilingual controlled vocabulary is generally not effective without extensive human intervention in the mapping process. This is due to the problems and idiosyncrasies of translation and usage of terms in various languages, which are not resolved with the simple employment of an automated dictionary or data mining.

5.4.1. Issues of Multilingual Terminology

The issues surrounding the development or implementation of multilingual terminology are discussed in detail in *ISO 5964:1985: Documentation—Guidelines for the Establishment and Development of Multilingual Thesauri.* In brief, issues related to mapping problems are listed below, ranked according to the difficulty of the solutions, from simplest to most complex.

Exact equivalence: The most desirable match involves terms in each language that are identical, or nearly identical, in meaning and scope of usage in each language. For example, the English *prayer nut* and the Italian *noce di preghiera* have the same meaning.

Inexact and partial equivalences: In cases where a suitable preferred term with the exact meaning and usage of the original term is not available in the second language, terms are sometimes linked as equivalents when they have only inexact or partial matches in scope and meaning. For example, the English *science* and the German *Wissenschaft* have overlapping but not identical meanings.

Single-to-multiple term equivalence: If there is no match in scope and meaning between terms, sometimes a concept in one vocabulary is matched to multiple descriptors in the second language. For example, the Spanish term *relojero* means both

watchmaker and *clockmaker* in English; however, in translation, the Spanish term could be repeated as a homograph and distinguished with the qualifiers *relojero (de pulsera)* and *relojero (de pared)* in order to map to the English terms.

Nonequivalence: Sometimes there is no exact match, no term in the second language has partial or inexact equivalence, and there is no combination of descriptors in the second language that would approximate a match. For example, the French term *trompe l'oeil* has no equivalent in English.

In the absence of an exact match between terms in different languages, inexact and partial equivalences may be used. Terms may be linked where both represent the same general concept, or where one term is broader and the second is narrower in meaning. When single-to-multiple term equivalences are made, a concept that is represented by a single preferred term in one language is represented by a combination of descriptors or a heading or phrase in the second language. In all of these cases, the definition or scope of the concept must be modified to cover the meanings of terms in all the languages.

None of the scenarios in the above paragraph is ideal. If the meanings of the terms differ significantly, it is better to fill a gap in one language with a loan term from the other language. A *loan* term is a foreign word or phrase that is routinely used instead of a translation of the term into the native language. For example, the term *lits à la romaine* refers to a particular type of bed peculiar to late-seventeenth-century French furniture; the best way to represent that term in an English language vocabulary is to use the French term as a loan term. Less desirable solutions include the adoption of a coined term in the second language. A *coined* term is a new term invented for the purpose of making a match between languages, generally by translating the term, but without authoritative literary warrant for the usage of the term. Terms without literary warrant should be avoided because they do not represent usage in the other language (and documenting usage is a critical criterion in creating terms); in addition, coined terms are often awkward at best and meaningless at worst. For example, if the French Gothic style term *Rayonnant* were translated into English as *Radiating,* it would be meaningless; the French term should be used in English.

If a new vocabulary is intentionally developed as a translation of an existing vocabulary, mapping between the two separate vocabularies is relatively easy. Mapping should occur from terms in an original language (called the *source* language) to terms in the second language (called the *target* language).

5.4.2. Dominant Languages

In a completely multilingual vocabulary, all languages are treated equally, with none serving as a so-called dominant language. However, in practical applications, it is often necessary to treat one language as the default dominant language, particularly when the vocabulary is rich and complex. An example is the *AAT,* in which each concept record includes over one hundred fields or data elements in addition to the term itself. With such vocabularies, it is impractical to maintain the data values of flags, notes, dates, hierarchies, and other subsidiary information in several languages. For the *AAT,* English is the dominant language, although terms and scope notes may be in multiple languages. In addition, if every term in the original source language has not been assigned equivalents in all other target languages, the status of the other languages is not equal to that of the source language, and they are known as *secondary* languages.

If a vocabulary such as the *AAT* is developed as a single unified vocabulary—but one in which the terms may exist in multiple languages—problems and issues with translations are resolved in the development process rather than in later mappings. Methods of development may entail the manual translation of the terms of the entire original vocabulary into another language or the addition of terms in several languages as each concept record is created. Creating such a vocabulary on the development side, rather than trying to map separate vocabularies later, makes the resulting set of multilingual terms very effective in searching across resources in different languages. In such a vocabulary, terms in different languages are exact equivalents, ideally linked only when meaning is synonymous and usage is identical or nearly identical. Issues of specificity and cultural context are taken into consideration in the selection of terms and the creation of relationships between concepts. Hierarchies and other relationships are likely to differ between comparable terminology in different languages, but such differences can be harmonized in development.

5.5. Satellite and Extension Vocabularies

Satellite and extension vocabularies may be considered *microcontrolled vocabularies* (also known as *microthesauri*), because they are specialized vocabularies that may fit into the structure of a larger, broader, or more generic controlled vocabulary.

A *satellite vocabulary* is characterized by having been constructed with the goal of being interoperable with an existing vocabulary. The satellite may be linked at multiple points to the original vocabulary. An example is a narrow specialty vocabulary that is intended to be integrated with the superstructure of a larger vocabulary.

An *extension vocabulary* is typically also constructed with the goal of being interoperable with an existing vocabulary, but is usually linked at one or a small number of nodes rather than being integrated at many points in the original vocabulary. *Node* or *leaf linking* is the method that links a specialized vocabulary to a node in the hierarchical structure of a broader controlled vocabulary so that the specialized vocabulary becomes a virtual new branch (or extension vocabulary) to the original vocabulary.

With either approach, the resulting family of controlled vocabularies should be consistent in structure, term format, and editorial oversight. By using satellite or extension vocabularies, specialized users may have access to the desired levels of specificity in the new controlled vocabulary without swamping the original controlled vocabulary with detail that may not be needed by most users. Furthermore, as noted in the discussion of local authorities in the following chapter, satellite and extension vocabularies can allow a particular set of users to access only the specialized vocabulary terms that apply to their indexing needs, thus excluding the full original vocabulary from these users, while ensuring that their specialized terms are still compatible with the full vocabulary in retrieval.