Design of Controlled Vocabularies

Elaine Svenonius
University of California, Los Angeles, Los Angeles, California, U.S.A.

INTRODUCTION

Language is the chief embodiment of information and it is, therefore, not surprising that in the information age, a new importance should attach to language. The 1980s have been called “the decade of language.”[1] Worldwide the present publication of language tools, such as dictionaries, glossaries, and thesauri, is proceeding at the rate of 500 per month and there now exist over 10,000 vocabulary standards in 35 languages.[2]

Controlled vocabularies (CV) form an important class of language tools. A vocabulary is a “list or collection of terms or codes available for use (as in an indexing system).”[3] A vocabulary is said to be controlled if it consists of a restricted subset of possible terms. Such a subset, in that it contains only those terms “authorized” for use, is sometimes called an authority list. In addition to terminological restriction, most CVs articulate semantic relationships between terms in the vocabulary, the most common of these being the inclusion of hierarchical relationship.[4]

Controlled vocabularies can take many forms and be used for many purposes. In the form of terminological databanks, CVs are used to assist in both manual and automatic translation. In the form of glossaries, they standardize and explicate the meaning or usage of terms in specialized fields of activity. In the form of literary thesauri they assist in composition by facilitating the expression of ideas. In the form of conceptual structures they give backbone to knowledge representation systems.

Within the discipline of library and information science, CVs are used primarily to assist in the retrieval of information. Classifications, thesauri, and subject heading lists are examples of CVs. If vocabulary is widely interpreted to include proper names, then name authority files are also examples of CVs in this domain.

The design of CVs within library and information science ranges from activities as simple as listing the allowable values for attributes in a specialized database to tasks as complicated as constructing an in-depth classification scheme. This article looks at the general principles underlying the design of CVs that provide alphabetic and classified approaches to subject indication. It begins with a statement of the purpose of CVs, then discusses the design of their vocabulary, syntax, and semantics. It concludes with a brief review of two issues of present concern: compatibility among CVs and effectiveness of CVs.

PURPOSE OF CONTROLLED VOCABULARIES

A major decision in the design of any information retrieval system is whether or not to incorporate a CV. Once the decision is made to adopt a CV, questions then arise as to how much and what kind of vocabulary control is to be admitted. Systems that do not employ vocabulary control may be characterized in terms of their indexing: natural language, derived, keyword, or title word indexing; or, in terms of the type of searching they allow: free-text searching or full-text searching. It is not a foregone conclusion that a system with vocabulary control is better than one without it. Indeed, as will be shown later, there are situations where simple keyword access to textual information may represent the most cost-effective alternative. Generally, however, it is assumed that CVs are essential in assisting users in their search for bibliographical information.[5] The assumption reflects the belief, widespread over time and place, that by controlling vocabulary it is possible to systematically correct some of the sloppiness in language that causes problems in retrieval.

In a perfect and orderly language, every object or concept would be designated by only one word, and each word would refer to only one concept or object. In such

---

[In this article terms will be used broadly to include verbal headings and class numbers.]

[The term controlled vocabulary is often used interchangeably with controlled language. There is imprecision in this usage insofar as a language, whether it be a natural language like English or an artificial language like a retrieval language, can be regarded as consisting not only of a vocabulary (i.e., set of terms), but also of a syntax and a semantics. On the other hand, the confusion is perhaps understandable since a controlled vocabulary incorporates the relational and referential semantics of a retrieval language as well as some of the syntax of the language.]

---

[Challenges to this assumption are reviewed in the last section of this article.]
a language there would be a one–one relationship, an isomorphism, between words and things. But neatness of this sort exists only in artificial languages. In any natural language, like English, there are synonyms (different words having the same referent) and homonyms (the same words having two referents). The consequence is a criss-cross of many–one and one–many mappings between words and things. Controlled vocabularies aim to make natural language a little more artificial. One way in which they do this is by straightening out these criss-cross mappings, that is by purging the language of synonyms and homonyms.

Synonyms and homonyms in a retrieval language are undesirable because they can cause failures when trying to communicate with an information resource, such as a bibliographic database. Synonyms can cause communication failures because a user searching on one of two synonymous terms, for example bulbous domes, might not find information that happened to be indicated by the other, for example onion domes. The chief requirement of any information retrieval system is to bring together all material on a given topic. This collocating requirement, or requirement for generic survey, cannot be met by a retrieval language infested with uncontrolled synonyms. A system that fails to meet the collocating requirement is considered to perform badly with respect to recall. Recall is a technical term defined as the ratio or percentage of all documents relevant to a search request that are actually retrieved. The various ways in which CVs address the synonym or recall problem will be discussed later in this article.

Homonyms in a retrieval language are responsible for a different kind of communication failure. A user seeking information on one of the meanings of the homonym drums, the musical instruments, may at the same time reference material on drums, the fishes of that name, containers for oil, or the components of a column. In other words, homonyms cause the retrieval of irrelevant material. To the extent that a retrieval system permits irrelevant material to be retrieved, it is considered to have poor precision capabilities. Precision, also a technical term, is defined as the ratio or percentage of material retrieved in response to a search that is relevant. The ways in which CVs address the precision problem will be discussed later.

The purpose of CVs, then, is to optimize the precision and recall capabilities of a retrieval language. Although precision and recall are fairly modern terms, use of the concepts has considerable historical precedent. In the middle of the last century, Samsom Low complained of the classifications then in use, that although they provided all the information in a given discipline, they failed to provide information on specific materials (quoted in Ref. [4]). In modern terms, he was complaining of bad precision at the expense of good recall. At the end of the century, Charles Ami Cutter “dealt the death blow”[5] to title term indexing, arguing that it failed to collocate related materials; in other words, it failed to provide for good recall. Cutter, in his famous statement of the objects of the catalog, first explicitly formulated what today is known as the recall requirement, and he came close as well to conceptualizing the precision requirement.[6] These requirements are reflected in Vickery’s demand that an indexing system provide generic survey and specific reference.[7] Perhaps their most succinct expression is Fairthorne’s stipulation that a retrieval system provide the user with all and only relevant documents.[8] Quite recently a new expression, representational predictability, has been introduced[9] to refer to the concept of recall. Precision and recall are the chief objectives of any retrieval language. It follows then, that any decision taken in the design of a CV must be rationalized with respect to improving either precision or recall.

VOCABULARY

Scope Definition

The first decision to be made in the design of a CV is which terms to include. This decision is governed first by a definition of scope and then by one or more selection principles. Scope definitions vary according to the type of CV being designed. They may be quite narrow, limiting their domain to the vocabulary of a single author, as is the case normally in back-of-the-book indexing; they may cover the vocabulary of a specific discipline, as is often the case with thesauri and special classifications; or they may be very broad, aiming to encompass most of the substantive vocabulary of a natural language, as in universal classifications and subject headings systems, such as the Dewey Decimal Classification (DDC) and the Library of Congress Subject Headings (LCSH). To be useful, a scope definition should address the specificity level at which the CV is to operate as well as its disciplinary breadth. Moreover, it should be formulated so precisely that candidate terms can be tested against it and either accepted for inclusion or rejected.

Selection Principles

Choice of terms to be included in a CV is more critical for CVs that index with verbal headings than for those that use classification numbers. Nevertheless classifications do include verbal headings and indexes and the principles enumerated below are, for the most part, applicable to them as well.
Literary warrant

Chief among the principles used to govern admission of a term into a CV is that of literary warrant. First introduced by Hulme as a means of class determination, the principle prescribes that CV usage be empirically derived from literature containing the vocabulary to be controlled. Thus, in back-of-the-book indexing, literary warrant sanctions the usage of the author of the book. For a periodical index the usage warranted is that represented in the core, and possibly, a few peripheral areas, covered by the index. Sometimes literary warrant is operationalized in terms of frequency of occurrence: terms that occur frequently in the core literature are candidates for the CV.

Use warrant

Vying in importance with literary warrant as a principle of term selection is use warrant, or common usage. Cutter called usage “the supreme arbiter” and since his time the principle of common usage has become well entrenched in the Anglo-American tradition of vocabulary design. A recent advocate of the principle is Dagobert Soergel, who in his blueprint for information retrieval systems design puts “request-oriented indexing” ahead of “entity-oriented indexing” in importance. Common usage has always been difficult to ascertain. The relatively new possibility of analyzing transaction logs of users’ searches on on-line databases offers some promise of progress in the understanding of terms commonly used in information seeking. Despite homage paid to it, the use warrant principle is frequently breached in CV design. Possibly some advantage is to be gained by including in a vocabulary terms a user would never think of, for instance, terms the user could be guided to for the purpose of improving precision or recall. But there is a real danger that such a practice can lead to a system that is both user unfriendly, by ignoring common usage, and unnecessarily expensive, by including terms that are never used.

Scholarly usage

Sometimes scholarly usage is favored over common usage or use warrant as a principle of term selection. Scholarly usage may be insisted upon when experts in a particular field are asked to assist in the construction of a thesaurus. Although experts are acclaimed masters of the language of their own fields and their opinion is valuable, there is a danger to relying on their opinion exclusively. What may be correct usage in the language of one expert may not be in that of another. And if students, for instance, rather than experts, are going to use the CV, some attention must be paid to their vocabulary. As noted, a thesaurus that ignores linguistic expectations can be both unfriendly and expensive.

The case is somewhat different for CVs that provide access through classification numbers. Here the vocabulary of users is not of primary concern. A classification scheme might well rely on scholarly terminology and benefit, thus, by its precision in the developing of taxonomies. The DDC uses scientific terminology frequently, for instance in its schedules for the zoological sciences.

Context Independence

In some CVs, particularly those used for indexing, it is desirable that terms be meaningful out of context. For instance, the term “ordinariness,” unless imbued with a specialized connotation as it is in the field of art, is not capable of standing alone as an indicator of subject content and, thus, is not suitable as an index term. Most thesaurus standards incorporate general rules aimed at discouraging the inclusion of terms that are not context independent, such as “Descriptors should be preferably in the form of a noun (or a noun phrase) or that form of verb which is grammatically equivalent” [International Standards Organization (ISO)]. These rules, not entirely successful in their aim, may be improved as more is learned about how the words of a particular field can be characterized into terms and nonterms.

The requirement that a term be context independent is not normally operative in classification schemes where the verbal headings are governed by the principle of hierarchical force. This principle says that what is true of general heading is true as well of the specific headings subsumed by it. Thus, in the DDC, Identification is an appropriate term at 371.952 because it is given context and meaning by the immediately superordinate heading 371.95 Gifted students.

Structural warrant

Another principle used in selecting terms for a CV is that of structural warrant. A term has structural warrant if it serves a useful collocating function within the CV. For instance, in a thesaurus, the term “masonry vaults” may be introduced though it has neither literary nor use warrant, but, as a supplied generic term, it collocates brick
vaults, stone vaults, and tile vaults and, in so doing, has
the potential for improving recall.

In classification schemes used to assign class numbers,
the need for structural warrant is met by encompassing
headings, for example, the DDC 344.071–344.072 Kinds
of Schools or by an enumeration of subsumed terms; 120
Knowledge, Cause, Purpose, Man.

A corollary to the principle of structural warrant is the
requirement that terms in a CV not be orphans, that is,
they have the property of connectivity in the CV syntactic
structure. A recent study of essential characteristics of
20 thesauri found that 11 of them contain no orphans.\[14\]
The requirement that there be no orphans in a CV is
not useful in every environment. If a CV is to be used in
conjunction with free-text searching, the requirement may
reasonably be used to cut costs; there is no point admitting
a term to a CV if it does not exhibit relations with other
terms, that is, if it does not have to be controlled. On the
other hand, if a CV is to be used in back-of-the-book
indexing, to eschew orphans would result in important
concepts not being indexed.

Consistency

Finally, a principle of consistency underlies term selec-
tion for a CV. If a CV contains the term “tests” and
several narrower terms, “group tests,” “intelligence
tests,” etc., the CV designer would probably not introduce
a term like “group intelligence testing,” no matter how
warranted by literary or patron use, choosing instead
“group intelligence tests.” Consistency considerations,
introduced often for the sake of structure, primarily affect
term form and frequently conflict with the dictates of
common usage.\[5\]

SYNTAX

Binding and Synthesizing Syntax

The word syntax derives from the Greek syntassein which
means “to put in order, arrange.” In reference to a natural
language, syntax means “the way in which words are put
together to form phrases, clauses, or sentences.”\[3\] In
reference to an index or classification language, syntax
refers to the arrangement of word forms to show their
mutual relations 1) in a term enumerated in a CV or 2) in
an index string that is formed by synthesizing enumerated
terms. In this article the former will be called binding
syntax and the latter synthesizing syntax. The assumption
will be made in this article that binding syntax alone
belongs to the domain of CV design, the domain of syn-
thesizing syntax being index language design. Thus, the
focus of this article will be on binding syntax; however, a
brief digression will be taken to clarify the distinction
between the two.\[6\]

The distinction between binding syntax and synthesiz-
ing syntax derives from another distinction: that be-
tween enumerative and synthetic index languages. In the
former all allowable expressions of the index language are
enumerated; in the latter some expressions are enume-
rated and some are synthesized by combining enume-
rated terms according to well-formulated syntax rules.
Probably the most significant development in index lan-
guage construction in the twentieth century is the move
from largely enumerative index languages to largely syn-
thetic ones.

Most index languages today, except those used in book
indexing, are synthetic in nature, consisting of enumer-
ated terms and rules for combining these into longer
expressions, sometimes called index strings. For example,
in the Library of Congress Subject Headings (LCSH),\[15\]
“columns, concrete” is an authorized main term enu-
merated in the schedules. Consisting of two words it is
called a compound term and the syntax which binds the
words is called an inverted syntax. LCSH permits the
use of the subordinate term “testing” under types of
engineering forms and, thus, “columns, concrete-testing”
is an allowable synthesized term in the LCSH retrieval
language. The permission rule is part of its synthesizing
syntax. Number building in the DDC and relating terms
using role operators in Preserved Context Indexing
System (PRECIS) are other examples of how elemental
terms in a CV are combined into longer expressions.
Synthesized expressions may assume a great variety of
different forms depending on the language in which they
figure. For an interesting review of various types of syn-
thesizing syntax, see Craven.\[16\]

\[1\]The two type of syntax, although conceptually distinct as the syntax
governing enumerated expressions and the syntax for constructing well-
formed synthetic expressions, are at times difficult to distinguish in
practice. In the DDC synthesizing syntax is expressed in the rules for
using the tables and the add instructions. In LCSH the synthesizing rules,
namely the rules for subdivision practice, take a variety of forms. A great
many of these rules are now incorporated in the LCSH manual.\[15\]
However, many of the permissible subdivisions used in LCSH are term
specific, in which case they are enumerated under the term in question.
Even some subdivisions of general application are enumerated in LCSH
even though, given the manual, this is redundant. In the PRECIS system
the two types of syntax are kept separate, the synthesizing syntax in the
PRECIS manual (1984) and the enumerative syntax in the rules that
govern the creation of the thesaurus used with PRECIS.

\[5\]For instance, common usage warrants “Vault shafts” but “Vaulting
capitals.”
Compound Terms

Two major decisions must be made in the design of enumerative or CV syntax. The first is the decision as to how many words to include in a CV term (the compound term problem) and the second is how to order the words in a compound term (the citation order problem). How many words to include in a CV term is at once the simplest and most difficult of decisions in CV design. Some standards (e.g., Ref. [12]) and some writers on the subject dismiss the problem with the observation that a descriptor refers to a concept and therefore should contain the number of words required to denote the concept. This approach, although it has the advantages of simplicity and being independent of any particular natural language, is not operational. Take, for instance, the two words “information” and “retrieval”: does their combination “information retrieval” refer to one concept or two, or many? Concepts are mental constructs with unclear boundaries and to use them as a basis for decision making invites inconsistency.

A more operational and less subjective approach to the compound term problem is that adopted in some automatic indexing programs,[17] viz. to collect frequency statistics for the occurrences of word pairs, triplets, etc., and then to stipulate that whichever of these occurs above a certain threshold frequency be admitted into the CV. This approach is strictly quantitative and as such does not need to rely on the notion of a concept; indeed one might even define concepts in terms of frequency data. Further recommending the frequency approach is its basis in literary warrant.

For thesauri designed to be used with on-line systems, it is generally considered desirable to restrict descriptors to a few words, if possible to only one word. Part of the rationale for Taube’s uniterm system, an indexing system consisting of one-word terms, was that no precoordinate syntax should be needed in a CV that could be manipulated with Boolean operators.[18] It proved not to be possible, however, to substitute postcoordinate syntax for precoordinate syntax for a variety of reasons: 1) most single words by themselves cannot pinpoint meaning and are thus unsuitable as index terms, 2) most single words do not provide enough specificity for the hierarchical structuring of a vocabulary, 3) most single words are not adequate for designating the major concepts of a discipline, for example the names of classes in scientific taxonomies, and 4) most single terms are often not effective as retrieval hooks in that they retrieve too many documents. Nevertheless, the terms in a CV designed for use in on-line systems tend to consist of fewer words than those in languages like LCSH, PRECIS, and other string index languages.

Most thesaurus standards provide guidelines for when two words, usually a modifier and a noun, should be kept together as a compound term. A consolidation of these would give the following: Retain a modifier and the noun it modifies as a compound term if: 1) the modifier has lost its original meaning (e.g., lawn tennis); 2) the modifier suggests a resemblance (e.g., tree structures); 3) the modifier does not form a subset of the class designated by the noun it modifies (e.g., paper tigers); 4) the modifier occurs frequently in conjunction with its noun (e.g., data processing); 5) the modifier is not widely distributed, (i.e., it is not used in conjunction with a great many terms); 6) the modifier is needed to show direction (e.g., fire engine or engine fire); or 7) the modifier is needed for collocation.

A specific guideline, promoted by Derek Austin[19] and reflected in the UNESCO[20] and British Standards Institution (BSI)[21] standards for thesaurus construction as well as in the PRECIS thesaurus, prescribes when not to retain two terms as a compound. Two terms should not be bound if 1) the modifier represents a whole and the noun a part or a property (e.g., aircraft engine), 2) the modifier is a patient and the noun is a transitive action (e.g., office management), or 3) the modifier is a performer and the noun is an intransitive action (e.g., bird migration). The relationship in each of the above instances is expressible in the PRECIS index language, that is, it is dealt with by synthesizing rather than binding syntax. The guideline is, thus, particularly, and perhaps only helpful if one is indexing with PRECIS. Its general applicability must await further investigation of the question when should two words be bound in a CV term and when synthesized in an index language expression.

An important factor to take into account in formulating rules for compound terms is the size of the CV. Too many compound terms can burgeon CVs and make them unnecessarily costly. Imagine a thesaurus, like an art thesaurus, in which nouns designating objects are modified by a variety of adjectives: adjectives designating style, material, technique, etc. One hundred objects each describable by any one of 10 adjectives would yield 1000 terms: wood bridges, wood tables, wood fences, etc. A limit must be put on this type of enumeration and normally this is done by replacing enumeration by synthesis.[22] In the LCSH, the Etc. device is an example of syntactic device used to limit the number of modifier–noun combinations: Mythology, Armenian (Dravidian, Egyptian, etc.).[8] A similar purpose is achieved in the DDC by the Add instruction. The more a CV admits of

--)
Design of Controlled Vocabularies

redundancy-reducing devices of this sort, the more synthetic it becomes—and, also, the more it assumes the unwanted properties of a uniterm CV.

In its study of the essential characteristics of thesauri, the Bureau Marcel Van Dyk\textsuperscript{14} measured the degree to which 20 thesauri admitted compound terms. The measure used was precoordination level and it was defined as:

\[
\frac{(a \times 1) + (b \times 2) + (c \times 3) + (d \times 4)}{a + b + c + d}
\]

where \(a\) = the number of uniterms, \(b\) = the number of two-word terms, \(c\) = the number of three-word terms, and \(d\) = the number of four-word terms. The scores for English, French, and German multilingual thesauri were, respectively, 1.7, 1.8, and 1.2; for monolingual thesauri, the scores were 1.9, 1.7, and 1.1.\textsuperscript{14} These statistics are interesting but should be regarded as normative only with caution: what ought to be cannot always be inferred from what happens to be the case. The better approach would be to perform research to investigate the effect of coordination level on retrieval performance. The main arguments for introducing compound terms into a CV are that they are needed for naming concepts, they give structure to the CV, and they help to narrow retrieved sets of documents, contributing, thus, to improved precision. The main argument for restricting the number of compound terms in a CV is that they burgeon the vocabulary. Not insignificantly they add to its cost: they cause a binding syntax to be implemented where often a synthesizing or even postcoordinate syntax could perform the same function at less expense.

Citation Order Problem

The second major syntax decision that needs to be addressed in the design of a CV concerns the order of words in a compound term. For compound terms consisting of one or more adjectives and a noun, the two major alternatives for citation order are to enter words in a natural word order or in an inverted word order. The first theoretician to address the problem was Cutter. Direct or natural word order was a fundamental principle of the dictionary catalog as Cutter envisioned it, because it enabled the man on the street to find what he wanted using familiar terminology.\textsuperscript{h} However, always willing to sacrifice consistency when it conflicted with the convenience of the user, Cutter provided an exception to direct word order in the case where another word in the compound term was decidedly more significant than the first word or was often used alone with the same meaning: "proper names" and "names."\textsuperscript{16} The lack of consistency fostered by a guideline that is based on a perception of significance, is all too apparent in the present day LCSH, which prescribes, for instance both "Jewish libraries" and "libraries, Catholic." The potential for inconsistency in Cutter’s rule led Swartz (quoted in Ref. [6]) and later Prevost\textsuperscript{23} to argue for a noun rule stipulating that in a compound term consisting of a modifier and a noun, the noun should always be in the first position. The effect of so rigid a rule, as Cutter noted, would be to put many subjects under terms where nobody would look for them: gastric juice under juice and alimentary canal under canal (see Ref. [23]).\textsuperscript{1}

Most thesaurus standards today advocate a direct or natural language word order syntax, suggesting, however, that inversions be entered as cross-referenced terms. Another method used in alphabetic subject systems to provide access to significant words in a compound term is to provide rotated terms displays; these provide keyword access to each word in the compound. In on-line systems, the citation order problem is not so serious, because access can be made to any word in a compound. There is still a decision, however, to be made about display: should all compound terms consisting of an adjective and a noun be displayed alphabetized on the adjective or might some classificatory sequences, utilizing inversions, be more helpful?\textsuperscript{j}

In the verbal headings enumerated in a classification scheme, there never were scruples about admitting inversions; indeed, they prove quite necessary. The advantage of inversion is that it not only often brings the most sought term to the fore, but it also provides some syntedic structure. For instance, the two terms "Egypt, ancient" and "Egypt, modern" collocate works on Egypt, which would be separated if the words in the terms were

\textsuperscript{1}A modern writer on the problem of significance and the order of words in compound terms is Eric J. Coates. Coates explicated significant to mean "conveying a definite mental image." Like Cutter he was resigned to the fact that significance order and natural language order were irreconcilable. His contribution to solving the problem was to suggest a categorization of words in a priority order of significance, for example, a word evoking a static image is more significant than one which expresses an action. Thus, in "Springing of cats," the more significant word is "Cat" (Ref. [24]).

\textsuperscript{2}Some book indexing standards, such as the British standard, concede that inversion may be used to bring the most significant term in the compound to the lead. It may be more appropriate in a CV designed for printed indexes to bend syntax to the task of achieving the recall objectives than in one designed for online retrieval.

\textsuperscript{h}Cutter’s staunch democratic stance was taken in opposition to those (elitists?) favoring a classed catalog, which required a user trying to find something on Badgers to understand a complicated scientific taxonomy (Ref. [4]).
entered in direct order. Cutter in elaborating on the syndetic structure for the alphabetic subject catalog sometimes worried that it might not have the collocating power of the hierarchical structures embodied in classification schemes.

**Prepositions and Conjunctions**

There are other syntax decisions to be made in the design of a CV. Generally these are of less moment than the two just discussed, namely, when to include compound terms in a CV and what order to use in arranging the words in the compound. Examples of such decisions are how to handle the use of prepositions and conjunctions in CV terms. Again different standards advise differently and, again, there are tradeoffs. Prepositions may be useful in promoting precision by resolving ambiguity or by juxtaposing, in Boolean-like fashion, two concepts: "artists as authors." On the other hand, they lead to longer terms, for example, "fertilization of flowers" instead of "flower fertilization?" Conjunctions may be useful both as precision and recall devices. The use of a conjunction between "religion" and "music," in the phrase "religion and music," serves, in a printed context, the same precision function as a Boolean **and** in an on-line context. (But how many conjunctions of this sort can a printed index support? and which ones?) Juxtaposing closely related terms by conjunctions, for example, "hotels and inns," may improve recall, but with a consequent impairment to precision, since a user interested in one must look and retrieve information under both.

In recent years interest in the syntax of CVs has grown. It is to be hoped that as more rigor is brought to the study of CV syntax, the syntax itself will become more formalized and better amenable to rationalization with respect to the precision and recall goals of CVs.

**SEMATICS**

Strictly speaking, **semantics** refers to the study of meaning; however, the word is loosely taken to refer to the different meaning structures found in languages. In a CV three such structures can be identified: 1) category semantics, 2) referential semantics, and 3) relational semantics.

**Category Semantics**

Normally the first structuring of a vocabulary of a subject field is a partitioning of it into semantically cohesive categories that are totally exhaustive and mutually exclusive. These categories, called facets, are classes of high generality. J. Kaiser, the first indexing theorist to make systematic use of facets recognized three facets: 1) concretes, 2) processes, and 3) country or locality. Ranganathan who developed and popularized faceted approaches in the construction of CVs, regarded the facets of each subject field as manifesting one of the five fundamental categories: 1) personality, 2) matter, 3) energy, 4) matter, 5) space and time. Vickery an early leader in faceting approaches to CV design in England, used eight rather explicit facets in his soil-science classification: 1) kinds of soil, 2) structure, 3) constituents, 4) properties, 5) processes, 6) operations, 7) laboratory techniques, and 8) general.

While facets can be regarded merely as highly general characteristics of division used in a CV, they can also be regarded as categories of existence. The postulating of categories or facets for a CV involves making an ontological commitment, a statement about what exists or is worthy of being indexed. Deciding what is worthy of being indexed, in other words, identifying the main concepts in an article, is often said to be the first step in indexing, while the second is to tool the concepts into the vocabulary of the index language. However, these two steps are not independent, since how the vocabulary is faceted predetermines, in Kantian fashion, what can be perceived as a main concept.

Many indexing theorists assume without question the value of incorporating faceting into a CV. But, as with other design considerations, advantages are offset by difficulties. Faceting is of particular use in CVs used with string index languages where they play a role analogous to the grammatical categories of a natural language. Just as the syntax rules of a natural language are defined in terms of well-formed sequences of grammatical categories, such as nouns, verbs, etc., so the rules for citation order (facet formulas) in a faceted index language are defined in terms of well-formed sequences of facet categories.

Although an index language may not use facets to perform a grammatical function, it might still realize some benefits by broadly categorizing its vocabulary into

---

Kaiser, wishing to give his index language a firm theoretical foundation, reasoned that all our perceptions are either of 1) things in general, real or imaginary, or 2) conditions attaching to things. Corresponding to these two modes of perception he recognized in his index language two categories or facets, Concretes and Processes. Vickery, in the tradition of English empiricism, argued that category definition should be based on literary warrant, that is, each subject should be allowed to yield up its own facets. Ranganathan, on the other hand, a rationalist par excellence, postulated a priori categories applicable to every subject field. The fact that things and processes, as well as time and place, appear in nearly all faceted index languages, making them index language universals, gives credence to Ranganathan’s position.
facets. Insofar as vocabularies that are faceted predetermine which concepts are deemed significant enough to index in an article, it would seem that indexing with faceted vocabularies would generally produce more consistent results than indexing with vocabularies that are not so categorized.

A more important use of facets is in the actual constructing of a CV. When collecting terms for a CV, it is convenient to divide them into manageable subgroups or facets. To require that these facets be semantically cohesive has the further advantage of simplifying, and ensuring correct order in the process of making equivalence, hierarchical, and related term relationships among terms in the CV. For instance, stipulating that equivalence and hierarchical relationships are to be constructed only among terms in a single facet can prevent the countless category mistakes made in hierarchical structuring, which relate terms from different facets, for example “traffic” and “traffic management” (a thing and a process). Stipulating that related-term relationships should not be made between terms in the same facet, because the facet structure provides sufficient relating, is a way of controlling the proliferation of related terms.

Helpful as facets may be in constructing CVs, their usefulness is compromised insofar as they are difficult to define. First, as was noted by Kaiser and later by Ranganathan, there are terms which cannot be categorized exclusively into only one facet. An example of such a term is “bibliography,” which Kaiser regarded as both a concrete (book) and a process (description) term. Another example is “fireproof,” regarded by Ranganathan as both a matter-property and a personality term. One means of dealing with multifaceted terms is semantic factoring, a technique recommended by Kaiser as well as some current thesaurus construction guidelines. In semantic factoring a concept is split into components, which when combined reproduce the original concept. For instance, a thermometer can be factored into device (a thing) + measurement (an operation) + heat (a property). A major problem with semantic factoring, however, is that it is not always possible to find factors for a concept, which when combined, give the original concept. For instance, “agriculture” consists of more than just “fields” + “cultivation,” that is, the whole is more than just the sum of its parts.

Another problem with category definition is the existence of terms, which by virtue of being ambiguous or vague resist categorization. An example is “organization,” which could be viewed as either a process or a thing. The problem here is easily fixed, however, the solution being to disambiguate the homonym in the CV, that is, to regard organization as two different words. (See the following section on Referential Semantics.) Posing a more serious problem to categorization and facet definition are terms that have no clear denotation. Terms referring to physical objects, like chairs, have clear denotations; one can point to a chair. But when language becomes unmoored from its groundings in existential reality and becomes abstract, the trouble begins. Kaiser and Ranganathan faltered at this point, as did Vickery. Particularly difficult to classify are mathematical terms such as coefficient and factor, and terms of energy such as labor, power, and light, and terms which refer to phenomena, such as love, truth, and beauty.1

Referential Semantics

We noted earlier that the chief characteristic distinguishing CVs from natural languages is their reference structure. To optimize precision and recall, CV designers attempt to achieve a homonym- and synonym-free language by setting up one–one relationships between terms and their referents. The handling of terms with more than one referent (homonyms) will be discussed in this section and that of synonyms, which involve both referential and relational structures, will be discussed in the following section.

The disambiguation of homonyms is handled differently in different CVs. In classifications schemes, terms that are homonyms are naturally disambiguated by context: they appear in different hierarchies and, thus, at different addresses. Often classification schemes have alphabetical indexes, which list the different or “relative” addresses where terms can be found. Thus, “mercury,” the element, may have addresses which locate it in the hierarchies for “metallurgy,” “chemistry,” and “mineralogy,” while Mercury, the planet, is located in the hierarchy for astronomy. A relative index not only disambiguates true homonyms: two or more words with

1One further difficulty in the definition of facets for CVs is caused by the fact that in index languages they perform both semantic and syntactic functions. Kaiser overlaid his semantic concrete/process distinction with a grammatical one: concretes are what are “spoken about” in a predicate and processes are what are “spoken of.” The distinction appears to be the subject/predicate distinction, but, anticipating tenents in modern linguistic theory, Kaiser warned that the subject/predicate distinction was a superficial one, whereas his was not.

Derek Austin probably more than any other indexing theorist sought to define facets grammatically. His facets include, among others, Key system (cf., what is talked about), Action, Object of Action, and Agent of Action. While these categories seem almost exclusively syntactic in nature, semantic and even morphological criteria are used in their definition. This at times leads to facet membership that appears counterintuitive, for instance the action category includes, in addition to verbs, phenomena terms like “Football,” “Disease,” and “Foreign relations,” because these appear to represent things in action. Defining facets along semantic dimensions is difficult enough; to require that the semantic distinction also be grammatical and/or morphological is ticklish indeed.
the same shape, for example ‘‘mercury,’’ but also distinguishes the various meanings of polysemes: one word with several meanings; ‘‘jackets,’’ which can be indoor or outdoor jackets.\textsuperscript{7} Another method of disambiguating homonyms, used in both thesauri and the alphabetical indexes to classification schemes, is by the use of parenthetical qualifiers, for example, mercury (element), mercury (planet), and Mercury (automobile). A homonym together with its parenthetical qualifier are considered to constitute a single compound term.

The degree of homonym disambiguation provided also varies among CVs. Often a classification or thesaurus is restricted to a particular domain of discourse. This limits the possible referents of their terms. Thus, there are character strings which are homonyms in natural language, (e.g., Mercury), but whose meaning is perfectly univocal within a particular CV (e.g., an astronomy thesaurus). Usually, such character strings do not need to be disambiguated by other methods, since their meaning is made univocal by the scope of the thesaurus. Further, if a CV is intended for use in an on-line rather than a manual environment there may be less need for presearch homonym control, since considerable disambiguation can be achieved by the postcoordination of terms.

Another device used to limit the meaning of terms in CVs is the scope note. Scope notes explicate how a term is to be used in the context of a given thesaurus or classification scheme. Normally, scope notes are not full-fledged definitions; however, there are cases when they must be expanded to such, for instance when a term is interpreted loosely in common usage or when it has various meanings that cannot be easily distinguished by a parenthetical qualifier.

It was remarked in the last section that terms that do not have well-defined referents are difficult to characterize.\textsuperscript{8} These terms also present obstacles to systematizing the reference structure of a CV. Among the terms in natural language that do have well-defined referents are many scientific terms, especially those that figure as variables in mathematical equations. No vagueness is permitted in the naming of concepts like mass, velocity, and acceleration. Two scientists will never disagree over the meaning of mass; it is not negotiable. The same cannot be said of two art historians discussing the meaning of style. Many of the concepts dealt with by humanists and social scientists have referents that are variable or indeterminate. Others cannot be neatly packaged into a word or short phrase, that is, they lack lexical expression (see Ref. [9]). Linguistic indeterminateness in the referential structure of a vocabulary presents serious impediments to imposing terminological control.

Relational Semantics

At the outset it was remarked that a CV was controlled in two senses: 1) its vocabulary was limited and 2) relationships among terms in the CV were made explicit. Control in the first sense is achieved by stipulating the referential structure of the CV by domain and facet definition and by devices such as scope notes, parenthetical qualifiers, and relative indexes. Control in the second sense is achieved by stipulating the relational semantic structure of the CV.

The relational structure of a CV is usually defined in terms of equivalence, hierarchical, and related term (RT) relationships. These are relationships of meaning: two terms stand in an equivalence relationship if they are equivalent in meaning, in a hierarchical relationship if the meaning of one is broader than the other, and in the related term relationships the two terms are related in meaning. It is implied by some thesaurus standards, for example UNESCO and BSI, that these meaning relationships are paradigmatic, that is, they are true a priori or by definition. As such they are contrasted with syntagmatic relationships, which are contingent and hold only in particular empirical contexts. An example of the distinction is that between ‘‘rats’’ which are always ‘‘rodents’’ (paradigmatic) and ‘‘rats,’’ which are sometimes ‘‘laboratory animals’’ (syntagmatic). While theoretically interesting, the distinction is not generally reflected in practice. Related term relationships particularly tend to be syntagmatic in nature and in many classification schemes, the hierarchical relationship is defined not according to logical principles, but rather to show context or perspective: ‘‘jackets’’ may be subsumed under ‘‘outdoor garments.’’

Equivalence

The equivalence relationship is probably the most valuable in any CV, because more than any other it serves
to bring together all material on a given topic; it is the chief means whereby the recall objective of the CV is achieved. The value of establishing equivalence relationships for a given knowledge discipline is directly proportional to the number of concepts in it that can be represented in a variety of ways. If the vocabulary of a discipline contains many synonymous expressions, it is said to be terminologically inconsistent [28] or to lack representation predictability. [9]

The mathematical or logical properties of the equivalence relationship are symmetry, reflexivity, and transitivity. Two or more terms standing in an equivalence relationship form an equivalence set. Normally, among terms in an equivalence set, one is chosen as preferred and see or use references are made from the others. But this is not the only way of handling equivalent terms. For instance, the singular and plural forms of a word, with some exceptions, are semantically equivalent from the point of view of information retrieval, but to make cross references from one form to the other would be wasteful. It is more cost effective simply to establish a general rule stipulating when singular and plural forms are to be used. An example of such a rule is the much–many rule, which distinguishes count nouns from mass nouns: if the question “how many” can be asked, for example, how many birds? the term is entered in the plural: if the question “how much” can be asked, as in how much hay?, the term is entered in singular. This rule is a simple one; more elaborate ones have been developed.

In traditional CV design, one of an equivalent set of terms is regarded as preferred and, thus, a standard terminology is established. Computer technology, however, permits a new kind of design wherein all equivalent terms are regarded as equal in retrieval. Such terms may be transparently linked in the computer so that a user sitting at a terminal and entering one of the terms is able to retrieve, automatically, all documents indexed by any of the terms in the equivalence set. Although a vocabulary standard is needed in a manual environment to prevent the scatter of related materials, vocabulary compatibility, which has the advantage of accommodating the language preferences of different groups of users, is a viable option in an automated environment. [9]

Another design question that arises in the handling of equivalence relationships is just exactly what terms are to be considered equivalent. Probably no one would quarrel that for the purposes of information retrieval, word form variants (cove vaults and coved vaults), syntactic variants (tables, mathematical and mathematical tables), and spelling variants (sulphur and sulfur) are identical in the sense of being intersubstitutable in all contexts. The problems in making equivalence determinations comes with synonyms. Actually, there are very few true synonyms in natural language. Even words that exhibit referential identity may have emotive overtones that affect their connotations: “guerrillas” and “freedom fighters.” From the point of view of information retrieval it might be assumed that two terms are equivalent if a user searching under one term would be as happy to retrieve anything that would be indexed by the other. From this point of view, emotive distinctions would probably not be significant, except maybe in some social science and humanities databases. Also probably not considered significant would be trade name variants (e.g., Kleenex and facial tissues), dialect variants (e.g., lift and elevator), or even scientific-popular variants (e.g., daisy and bellis perennis). It is a different question, however, when the terms to be linked are near synonyms, terms whose meanings overlap widely (e.g., marshes, bogs, wetlands, swamps, and marshlands). Regarding these terms as equivalent does affect retrieval, improving recall for users to whom the fine distinctions between the terms do not matter, impairing precision for those to whom the distinctions do matter. Whenever there is an apparent trade-off between precision and recall in a CV the designer has the choice of either favoring one or the other, based on knowledge of his users, or attempting to find ways to accommodate both, for example through efforts to achieve vocabulary compatibility (q.v.).

In some CVs the see or use reference is broadly conceived as a line between terms that are not used, but form part of a lead-in vocabulary, to terms that are used. Thus, it can happen that these references link not only terms that are roughly equivalent but also those with broader and narrower meanings. Sometimes specific to general see references are used, for example, “plant waxes see waxes,” or “automatic transmission fluids use automatic transmissions and transmission fluids,” in order to control the specificity of the CV and, thereby, the number of terms in it. Occasionally see or use references are made from broad terms to more specific ones: “asylum see political asylum,” the justification being that the broader term is always used in the narrower sense. Sometimes antonyms are linked by see or use references: “hardness” and “softness,” the rationale here being that the terms represent opposite points on a continuous scale and, thus, really refer to the same concept. The wisdom of broadly interpreting use and see references is debatable, especially given the possibility of transparent linkages,

Footnotes:
[9] An example of an exception is when the process-thing distinction is sometimes reflected in the singular–plural forms of a term, for instance, “Painting” (the act of) and “Paintings” (the objects).
[8] This is true at least as far as retrieval is concerned; in some contexts, the display of retrieved items requires selecting one of an equivalent set of terms as preferred. (See section on “Vocabulary Compatibility”).
and needs empirical testing with reference to the precision and recall needs of the users of the CV.

Hierarchy

Almost all CVs incorporate classificatory structures embodying hierarchical relationships in one form or another. Like the synonym relationship, the hierarchical relationship serves chiefly to improve recall. A given concept can be referred to by different terms and while, normally, this is understood to be the situation of synonymy, it also characterizes hierarchy, where one concept is referred to by more than one term but these terms are expressed at different levels of generality. How a particular concept is represented in a retrieval system is difficult for a user to predict. Studies have shown that users of an information retrieval system tend to formulate their search terms either at a higher or lower generality level than that employed by the system (e.g., Ref. [29]). Thus, to achieve good recall it is necessary to bring into coincidence the user’s vocabulary and that of the system CV.

Hierarchical structures can also be used to improve precision. For instance, they can assist users who have anomalous information needs, or who are simply inarticulate. An area of a classified display may be likened to a ballpark into which a user either strays or is guided by an index and serendipitously finds what he wants by browsing. In online systems with Boolean capabilities, hierarchical structures can improve precision, in another way, viz. by eliminating large subsets of unwanted documents. Imagine search on “protein and women,” which may retrieve upward of 1000 documents. A quick perusal shows that many of these documents deal with protein requirements of women suffering some ailment, whereas what the searcher intended was material on protein requirements of healthy women. The retrieved document set can be reduced dramatically by qualifying the search prescription by “and not pathological states.” To use such a qualifier, however, presumes hierarchical relationships have been established between “pneumonia,” “scarlet fever,” etc. and the term “pathological states.”

The hierarchical relation can be variously interpreted. In its strictest interpretation its prototype is the genus–species relation, also known as the inclusion relation. This relation has the mathematical properties of reflexivity, transitivity, and antisymmetry. It also has the mathematical property of inheritance, in the classification literature called hierarchical force, whereby whatever is true of a given class (animals) is also true of all classes subsumed by it (e.g., warthogs, deer, gazelles, etc.).

Many of the hierarchical relationships in traditional classifications and thesauri are inclusion relationships in the sense just defined, but a great many are not. Fairthorne gives the excruciating example from the Universal Decimal Classification (UDC), wherein “slide rules” are subordinated to “calculating and adding apparatus,” which in turn are subordinated to “wheel mechanisms.” While this juxtapositioning may simply be a mistake, there are many instances in both thesauri and traditional classifications where hierarchical structuring to show relationships other than that of strict inclusion is quite deliberate. For example, some thesaurus guidelines sanction whole–part, topic–subtopic, and geographic region–subregion as legitimate hierarchical relationships. Many book classifications recognize point-of-view relationships as hierarchical. For example, the DDC does not, in genus–species fashion, subordinate “insects” to “invertebrates,” but rather to “disease carriers.” While not all insects are disease carriers, it has been deemed helpful to create a path in the classificatory hierarchy that looks at them from that point of view.

The decision of how to interpret the hierarchical relationship in the design of a particular CV depends on a variety of factors, chiefly the nature of the vocabulary to be structured and the purpose for which it is structured. If a vocabulary includes many terms whose referents are linguistically indeterminate, it makes little sense to insist that the relationship be defined in terms of strict inclusion. Indeed, as Plato experienced, even to ask what inclusion relationships a term like beauty partakes in is to fall into a linguistic quagmire.

The hierarchical relationship is interpreted differently according to the purpose of the CV for which it is intended. Classification schemes designed to aid users in seeking the right book on the shelf may favor point of view or perspective hierarchies over the more strictly defined structures of thesauri, whose propose is to lead the user seeking the right term. Other structural differences between book classification schemes and information retrieval thesauri, attributable to differences in purpose, include differences in 1) connectedness, 2) types of terms admissible, and 3) class and subclass formation.

Most classification schemes attempt to map a domain of knowledge and are, thus, obliged to show the location of each subject in the scheme relative to all others. Such a

---

*If CVs are to be used for the purpose of knowledge representations in expert systems, it is essential that their hierarchical relationships have the property of inheritance, that they be defined strictly in terms of the inclusion relation. This property is assumed in the syllogistic deductive reasoning used by expert systems.*
system of relationships possesses the property of connectedness. This property is not needed in thesauri. While the classificatory structure of a classification like the DDC is often likened to a gigantic upside down tree, that of a thesaurus might be said to resemble a collection of small shrubs.

Terms to be admitted into a thesaurus intended for indexing use must be capable of conveying meaning in a context-free environment. As was noted above, this is not a requirement for terms in a classification scheme, part of whose meaning is acquired from the hierarchical contexts in which they appear. For instance, a term like ‘‘other groups’’ (meaning ‘‘other kinds of oxides’’) can appear in a classification scheme but never in a thesaurus. Even more to the point is that such terms are never used as retrieval hooks, only the numbers associated with them are. Thus, the hierarchical relationships admitted in a thesaurus, being limited to relationships between context-independent terms, are much more limited than those in a classification scheme.

Certain canons or principles have guided class and subclass formation in book classification schemes to ensure that no book is ever put in two classes and that every book is put in some class. One of these is the canon of mutual exclusivity. First articulated by Aristotle to ensure there would be no cross-classification in nature, it requires that all the subclasses of a class be nonoverlapping. It requires as well that a given class have only one broader class. This canon, while essential in classification schemes designed for arranging books along shelves, is not a requirement for thesauri, which can be configured as lattices as well as trees, that is, they admit of partial orderings as well as linear orderings.

Related term relationships

The related term (RT) relationship serves to stimulate the verbal imagination of the user of a CV thereby leading him to terms more appropriate to his search topic than those originally coming to mind; it has, thus, both recall and precision functions. In traditional and many present day CV guidelines, the RT relationship is vaguely defined. It includes all semantic relationships, other than those of equivalence and hierarchy, and the only mathematical property it always possesses is that of symmetry: if A is related to B, then B is related to A. In the ANSI standard two terms are said to stand in the related term relationship ‘‘if it is believed that the user, when examining one of them, might want to be reminded of the existence of the other.’’ Further definition is given in the form of enumerated examples, for example, terms are RT related if they are nearly synonymous, have viewpoint interrelationships, or represent concepts that either overlap or bear a whole–part relationship to each other.

Perhaps the least restrictive condition ever imposed on the RT relationship was one used in the early days of LCSH wherein two terms were related if the concepts they referred to occurred in the same monograph.

The general lack of rigor in expressing conditions for RT relatedness may be tolerable on the grounds that even a scatter-shot approach to linking terms can improve retrieval effectiveness by helping the user to be more articulate. However, opinion seems to be mounting that more control should be exerted over RT relationships in order to avoid inconsistencies and subjective judgments. For instance, a guideline suggested by the UNESCO standard for thesaurus construction is that of two related terms, one should be strongly implied by the other. More specifically, two terms are related if one is a necessary component in the explanation of the other, as for example ‘‘birds’’ is used in the explanation of ‘‘ornithology.’’ The guideline could be made even more operational by specifying that the explanations to be used be taken from certain authoritative reference works. However, to restrict the RT relationship to only those terms that are definitively related, while in accordance with the view that a CV should exhibit only paradigmatic relationships is to improve precision and recall by helping users to formulate better search strategies.

Another approach to operationalizing the RT relationship is to specify, in the form of schema, the particular relations it may encompass. Examples of such relations that might be considered are: ‘‘is presupposed by,’’ ‘‘causes,’’ ‘‘is made of,’’ ‘‘is used for,’’ ‘‘is produced by’’ (see Ref. [31]). By examining already existing CVs in a given field, one might try to develop a set of relations tailored to that field. A difficulty with this approach, however, is that often two terms are helpfully related in ways difficult to verbalize, in other words, the relationships are lexically indeterminate.

An economically attractive solution to the related term problem is to eschew making RT relationships altogether. This approach is particularly worth considering in the online environment where other means exist for suggesting vocabulary to users. For instance, computer algorithms could be written to present to the user at the time of his search the various terms that have been associated with his terms in past searches. Potentially the number of terms associated with any given term is very large, but actually the number frequently associated in retrieval

---

*Personal communication from Mary K. Pietris, Chief, Subject Cataloging Division, Library of Congress.

†Some online systems keep a record of terms correlated in search strategies in the form of Boolean disjunctions. Terms that are frequently ORed together are made available to users of the systems to help them in formulating their search strategies (Ref. [32]).
must be rather small. Besides being economically attractive using co-occurrence data from user requests to operationalize the RT relationship realizes the UNESCO and ISO strictures that related terms should be established only if they would be required in retrieval.

**Degree of Control**

An important design consideration relating to the semantic structure of a CV is how much control to impose. A routine adherence to present guidelines for the construction of classifications and thesauri can blind one to the fact that there is a choice in this matter. It is, moreover, a choice with financial implications, since the cost of a thesaurus is a function of the number of semantic relationships it recognizes.

Degrees of control correspond to degrees of vocabulary normalization. A very minimal level of control is to recognize as equivalent the orthographic variants of a term: "database," "data-base," and "data base" or "color" and "colour." Also at a fairly minimal level is the control achieved by the normalization of singular and plural variants. A much greater degree of control comes with the introduction of semantic equivalence relationships and a still greater degree with the establishing of hierarchical relationships. Maximum control is achieved with the recognition of the various relationships embedded in the RT relationship. Possible decisions that might be considered in thesaurus design are to limit control to the level of semantic equivalence or to not admit RT relationships.

The Bureau Marce van Kijk has developed measures to describe the amount of control exhibited in different thesauri. Various thesauri were sampled to determine the number of their hierarchical (ascending and descending) and related term relationships. For 10 monolingual thesauri examined, averaging 2850 terms, the mean number of hierarchical relationships was 4900 and that of related term relationships was 7200. Another control measure was richness of structure (or accessibility), defined as the ratio of the number of semantic relationships in a thesaurus and the total number of terms contained in it. For the 10 monolingual thesauri the richness of structure ratio was 4.79.

In making decisions about the amount of vocabulary control needed in a given situation, consideration might be given to how much control can be effected automatically and how much can be ignored and relegated to users. It is reasonable to expect that computer algorithms could be developed to equate most orthographic and singular–plural variations; and, as was suggested earlier, algorithms could be developed to derive related term relationships. The use of algorithmic methods to obtain vocabulary control is attractive as it promises the reduction of intellectual effort and, therefore, cost.

A CV need not take the full burden of vocabulary control required to conduct a successful search. Certain aspects of it could be left to the user to perform at the time of search. It might make sense, for instance, to let users define, using Boolean disjunctions, those relationships that are viewpoint dependent. User-imposed vocabulary control, however, is only as successful as users are articulate in formulating their searches.

**VOCABULARY COMPATIBILITY**

**History**

The proliferation of tools to effect vocabulary control has had the unfortunate, but perhaps not unexpected consequence, that now new tools are needed to control the CVs. The problem that needs to be addressed is that of a user attempting to retrieve information from multiple databases for one search topic, each governed by a different CV or that of two indexing agencies serving similar users, wishing to consolidate by developing a common CV. In the first instance it would be ideal if the user's search prescription, formulated in his own words, could be automatically translated into the CVs of the different databases; and, in the second, if the two indexing agencies could merge their vocabularies.

The possibility of developing automatic methods to render two CVs compatible was raised in the early 1960s. At that time two different but related terms were coined, the first descriptive of vocabularies and the second of systems: 1) Convertibility, defined as "the ability of one indexing vocabulary to another;"[34] and 2) Compatibility, defined as "the ability of one information system to accept the original indexing and abstracting data of another information system for any given subject coverage that is common to both systems."[35]

Early attempts to achieve compatibility took the form of directly translating one CV into another. This method, which is still used, requires the construction of a conversion table, an example of which is the Dictionary of Equivalents developed to link the vocabularies of the Armed Services Technical Information Agency and the Atomic Energy Commission.[36] It was perceived by some that efficiency might be gained if instead of providing translations between individual pairs of CVs, a meta or switching language were developed that could take as input a term from one CV and, through its mediating
structure, translate the term into a number of other CVs. The Intermediate Lexicon, developed by the Groupe d'Étude sur l'Information Scientifique in Marseilles, was the first, so called, switching language.\[37\] Perhaps the best known switching language is the Broad System of Ordering, developed by UNISIST for the purpose of linking different classifications and thesauri in information retrieval.\[38\]

In Europe today, where language barriers abound, practical attempts to achieve CV compatibility focus on the construction of multilingual thesauri, an example of which is Eurodicautom, a system designed for users of Euronet.\[39\] On this side of the Atlantic interest in compatibility is largely confined to facilitating retrieval where there are a number of English language databases on a single subject. A well-known example is the R.T. Niehoff project at Battelle Laboratories, the objective of which was "to identify, compile and integrate existing energy vocabularies from systems, both government and nongovernment, into a common indexing and retrieval guide.\[40\]

### Design Problems

Strong compatibility between two CVs, in the sense that there is perfect translatability between them, is not achievable. The reasons for this are in part lexical and in part due to the relational semantics, particularly the classificatory structure, of the two CVs. Perhaps the most usual type of incompatibility occurs when two CVs operate at different levels of generality ("poverty areas" and "skid rows"). But there are other sources of incompatibility as well. It may happen that a term exists in one CV, but not in another. Or the same term may exist in the two CVs but with slightly different referential or point-of-view meanings. Two terms might be differently contextualized in the two CVs because partitioning is done by different characteristics of division or hierarchical relationships are defined by different rules. As a consequence of these problems, there will always be some information loss when translating from one CV to another.\[41\]

Although strong compatibility between two CVs is not possible, a form of weak compatibility can be achieved in practice. Two CVs can be used together (jointly) by merging them and attempting to create a semantic structure that accommodates both. A model for how this is done is provided by the on-line user who searches a multidatabase system with a Boolean disjunction of terms from several CVs. Precoordinately merged vocabularies designed to help such a user have existed in manual environments for some time, an example being the Cross Reference Index: a Subject Heading Guide, which clusters related concepts in the vocabularies of LCSH, Sears' Subject Headings, Public Affairs Information Service, the New York Times, and the Business Periodical Index. In the medical field an example of a merged vocabulary is the Guide to LC/MeSH Equivalents in Health Science Books. Among the merged vocabularies used in on-line systems is the Social Science and Business Microthesaurus, which is used to search the National Technical Information Service (NTIS) database and integrates the vocabulary and hierarchical structures of five CVs. Another is TERM, used in searching the Bibliographic Retrieval System (BRS) database and integrating social science subject headings from five thesauri as well as free text terms related to the subject headings. An example of a very large merged vocabulary currently under development is the Art and Architecture Thesaurus, which incorporates the vocabulary of several different art and architecture indexes as well as appropriate vocabulary from LCSH.\[42\]

Among the tasks to be faced by designers of CVs in the future is the development of standards or guidelines for the construction of merged vocabularies. Research is needed to determine where merging can be done automatically and where it must be subject to human review. Such research might prove useful as well in revising present thesaurus construction standards, which seem to be slow in adapting to the age of on-line retrieval, where CV searching is done in conjunction with free text searching and where multiple database searching is becoming the norm. It has been hypothesized that the problem of CV proliferation has led on-line database users to turn increasingly toward free text searching. Unless merged vocabularies, perhaps in the form of metathesauri, are developed to help such users, the future of some database-specific CVs may be in jeopardy.

### Effectiveness of Controlled Vocabularies

A CV can improve precision and recall in retrieval, yet the expense of creating them and the inconvenience of using them has given rise to questions of their effectiveness. This section will review briefly historical positions taken on the value of CVs and then discuss factors to be considered in deciding whether to construct a CV for a specific situation.

\[This hypothesis was suggested by Raya Fidel and appears to be receiving support from data she is analyzing (private communication).]
History

As noted earlier, the use of uncontrolled vocabularies, the alternative to CVs, takes two forms: to index with keywords—significant words taken from the titles and abstracts of documents—or to not index at all but to provide free-text searching of the subject rich parts of document records or the documents themselves. The issue of uncontrolled vocabularies versus CVs has a history that can be divided into three eras. The first began in the last century when title term indexing, or title catchword indexing, a precursor of the modern keyword indexing, was proposed by Samson Low and his assistant Andrea Crestadoro. Providing access to books by words in their titles was regarded as a solution to the precision problem caused by the classed catalog. Cutter was opposed to title term indexing for two reasons: 1) titles might not express the true subjects of the works they name and 2) works on like subjects would be separated if different words were used in their titles, for example, "free trade," "protection," or "tariff." Cutter is credited with having dealt the deathblow to title term indexing.

His Rules for a Dictionary Catalog (1904) provided guidelines for the construction of alphabetic CVs that are as valid as, and in some parts even more sophisticated than, those in use today. In any case, his arguments against uncontrolled vocabularies seemingly won the day: the LCSH based on Cutter’s rules, became the predominant means of subject access in libraries, and controversy over the effectiveness of CVs subsided for half a century.

The second era in the controlled versus uncontrolled vocabulary controversy came with the advent of the computer and the promise it offered of derived, or natural language, indexing. The best known of the derived indexing methods, Keyword in Context Indexing (KWIC), was introduced by Hans Peter Luhn in 1959. Given the promise it offered of quick and economical indexing, the question naturally arose, is vocabulary control really necessary? The question was addressed in the early 1960s in England in an experiment dubbed Cranfield II. The purpose of the experiment was to compare the retrieval effectiveness, measured in terms of precision and recall of 33 indexing languages incorporating various degrees of control. Its most widely cited result was that a minimally controlled vocabulary, one in which only synonyms and word endings were normalized, gave results as good as, and sometimes even better than, full vocabulary control. A series of similar experiments, reapplying the Cranfield methodology followed; they produced comparable findings and added grist to the mill of those who doubted the value of vocabulary control.

The Cranfield II findings were widely accepted, despite flaws in the methodology of the experiment. It was not until the mid 1970s that they were challenged by experiments of a different and more modest kind. In 1976 Barbara Charton searched the Chemical Abstracts keyword index, an uncontrolled vocabulary constructed of keywords selected from the titles and texts of documents, for information on "correlation analysis." She then examined the actual issues of the journal for information on this subject and discovered that the keyword index failed to indicate 50% of the articles she found. She complained to the editor of Chemical Abstracts, who responded that the set of keywords she had used to search the keyword index were not complete and, furthermore, the keyword index was designed for current and quick subject access; it was not the tool to use if her aim was good recall. He then referred to a study, performed by him and his colleagues, which showed that "searching controlled and uncontrolled vocabulary files...gives complementary but not necessarily identical results."

Era three in the controversy over controlled versus uncontrolled vocabularies is the present, the era of on-line searching. The prevalence and popularity of free text searching again has led to the value of CVs being challenged. Various experiments similar to Charton’s, but in the context of on-line databases have been conducted. One of the earliest was carried out in 1977 by Carrow and Nugent. They compared free text and controlled vocabulary searching of the National Criminal Justice Reference database. They discovered that the two methods had about the same precision, but controlled vocabulary searching produced significantly better recall. Their conclusion was like that of the editor of Chemical Abstracts but formulated as a hypothesis, viz., the best performance would be achieved by a combination of the two methods; they were complementary. A number of studies of a similar vein produced similar conclusions, perhaps best articulated by Hend.

---

*A not unusual criticism of any experiment is that results emanating from a laboratory situation, where certain variables are controlled, do not accurately reflect reality. In the Cranfield experiments, this type of criticism was aimed at the sets of questions used to represent users’ information-seeking behavior and the search strategies used to retrieve documents. Also criticized were the samples, the definitions of dependent and independent variables, and the statistics employed in the experiments.*
zler: ‘‘Thus, the alternative ‘free-text or controlled vocabulary’ is no longer an alternative: there should always be both free-text and controlled vocabulary in an ‘ideal’ combination.’’[53]

Factors Affecting the Decision to Construct a CV

Whether a CV should be constructed in a given situation depends on a number of factors. Some of these are obvious, such as existence of closely related CV, the availability of financial and intellectual resources, and the political promise of support. There are others, however, which are not so obvious, that have to do with the expected effectiveness of a CV. Foremost among these are the nature of the subject discipline involved and the retrieval requirements of potential users of the CV. If a subject discipline is such that its writers tend to give their works noninformative titles, a CV is needed. A CV is needed if the vocabulary of the discipline exhibits very little representational predictability. On the other hand, if the discipline lacks a special terminology or if its vocabulary exhibits a great deal of linguistic indeterminateness, a CV may be less valuable to the extent that control is difficult to impose. It may be less valuable as well in the discipline where the predominant mode of organizing and searching for information is not by subject (in disciplines whose information resources are largely archival).

CONCLUSION

Users’ retrieval requirements vary from situation to situation. There are situations where the effects of vocabulary control might not go unnoticed for the most part, for instance, where the user’s goal is current awareness browsing, rather than retrospective searching; or where a user is interested in only one or two documents on a subject and does not need the collocation afforded by a CV. On the other hand, there are situations where such collocation cannot reasonably be sacrificed, for instance, for the scientific researcher who searches a database to see if his discovery is a new one. There are also situations where the precision afforded by a CV may be essential, for instance to reduce a large set of documents produced by an online free text search that uses terms widely and frequently distributed in the database. Perhaps as near as one can come to generalization about the value of a CV is simply to say where precision and recall are important retrieval objectives, then a CV of some kind is mandated.

ACKNOWLEDGMENT

I would like to thank Martin Dillon, Director, Office of Research, OCLC for his suggestions.

REFERENCES


40. Niehoff, R. The Design and Evaluation of a Vocabulary Switching System for Use in Multi-Base Search Environments; Battelle Columbus Laboratories: Columbus, OH, 1980.


