Government of Canada Gouvernement du Canada

Guide to the Development and Maintenance of Controlled Vocabularies in the Government of Canada

Prepared by the Controlled Vocabularies Sub-Group of the Government On-Line (GOL) Metadata Working Group

July 8, 2005

Canada

Catalogue No.
ISBN

This document is available on the Treasury Board of Canada Secretariat
Web site at www.tbs-sct.gc.ca

This document is also available in alternate formats on request.

Table of Contents

Acknowledgements

This guide is a project of the Controlled Vocabularies Sub-Group of the Government On-Line (GOL) Metadata Working Group. It draws on the expertise of information management professionals in several disciplines and government departments who pooled their thoughts on the issues and best practices involved in the application and development of controlled vocabularies.

For their support in the creation of this guide, the Sub-Group thanks:

Name	Department
Kate Carter	Industry Canada
Marie-Claude Côté	Treasury Board of Canada Secretariat
Mimi Golding	Transport Canada
Patricia Gorman	Social Development Canada
Elizabeth Kirby	Public Works and Government Services Canada
Don Knorr	Canada Revenue Agency
Gabriel Lepkey	Health Canada
Gay Lepkey (Editor)	Publishing and Depository Services, Public Works and Government Services Canada
David L. McCallum	Facilitator/Writer
Karen Morgenroth	Canada Institute for Scientific and Technical Information, National Research Council Canada
Anne Price	Library and Archives Canada
Gail Rawlings	Office of the Auditor General of Canada
Cecil Somerton	Treasury Board of Canada Secretariat
Freda Taylor-Christopher	Statistics Canada
Alain Vaillancourt	Public Health Agency of Canada

Chapter 1: Introduction

This guide is intended for the use of information professionals within the Government of Canada (GC) who are faced with the challenges of adopting or adapting existing controlled vocabularies or creating entirely new controlled vocabularies that will provide values for GC metadata elements.

The guide was developed by members of the Controlled Vocabularies Sub-Group (see http://www.tbs-sct.gc.ca/im-gi/mwg-gtm/cvsg-sgvc/intro_e.asp) of the Government On-Line (GOL) Metadata Working Group (see http://www.tbs-sct.gc.ca/im-gi/mwg-gtm/intro_e.asp).

The guide identifies and explains the various types of controlled vocabularies and their characteristics. It outlines the advantages and challenges they represent and presents a systematic approach to evaluating them for use. It also provides references to related resource material. It is not designed to serve as a comprehensive text on a subject of considerable complexity; however, it is hoped that the guide will provide useful information for planning and resource allocation purposes, including the development of appropriate business cases.

The construction of controlled vocabularies requires that those involved become familiar with a range of additional resources and standards. Large and complex controlled vocabularies require the involvement of personnel with considerable training and experience, while smaller ones require less expertise.

The guide addresses controlled vocabularies in the context of requirements for creating and maintaining the metadata elements mandated by the Common Look and Feel (CLF) for the Internet, Standard 6.3 (see http://www.tbs-sct.gc.ca/clf-nsi/inter/inter-06-03_e.asp). This standard grew out of a government-wide directive issued in 2000 by the Treasury Board of Canada Secretariat making it mandatory for GC Web sites to include descriptive metadata, i.e. structured information about the content of Web-based resources. A Web resource is defined as a single Web page, a document (consisting of multiple Web pages), a digitized image, a sound file, or an animation such as a movie.

Comments on this guide and suggestions for improvements are welcome and should be directed to the attention of the Chair of the Controlled Vocabularies Sub-Group by e-mail at meta_coord@lac-bac.gc.ca.

1.1 Metadata in the Government of Canada Context

The *GC Metadata Framework* (http://www.tbs-sct.gc.ca/im-gi/meta/frame-cadre_e.asp) establishes an overall strategy for the development of metadata within the Government of Canada.

With TBITS 39.1 (http://www.tbs-sct.gc.ca/its-nit/standards/tbits39/crit391_e.asp), the Treasury Board adopted the international Dublin Core metadata standard for use in the GC. Of the 16 Dublin Core metadata elements, five required by CLF Standard 6.3 are mandatory; the rest are optional. For information on the Dublin Core Metadata Initiative, refer to the related Web site at http://dublincore.org/.

For guidance on applying metadata in the GC context, please see the *Government of Canada Metadata Implementation Guide For Web Resources* (4th edition, October 2005) at http://www.tbs-sct.gc.ca/im-gi/mwg-gtm/ts-sf/docs/2005/migwr-gpmrw/migwr-gpmrw00_e.asp . The *Metadata Implementation Guide* is a useful starting point for those with little or no knowledge of metadata or its role within the GC.

Some departments or agencies employ more metadata elements than mandated by CLF Standard 6.3. This guide may also be used to determine appropriate controlled vocabularies for those elements.

For the mandatory element dc.subject, TBITS 39.2 (http://www.tbs-sct.gc.ca/its-nit/standards/tbits39/crit392_e.asp) mandates the *Government of Canada Core Subject Thesaurus* (http://www.thesaurus.gc.ca), or CST, as the default controlled vocabulary to describe the subject(s) of GC Web resources. This means that, in the absence of any other appropriate registered subject vocabulary, the CST must be used as a source of vocabulary for dc.subject.

1.2 What Is a Controlled Vocabulary?

The Controlled Vocabulary Sub-Group has adopted the following definition of a Controlled Vocabulary:

> [A controlled vocabulary is] a list of standardized terminology, words, or phrases used for indexing or content analysis and information retrieval usually in a defined information domain. It is characterized by consistent format and syntax and may include synonyms and cross-references. In a controlled vocabulary, one of a set of possible terms representing a concept can be used as the representative term for that concept. Consequently, all resources about, or pertinent to, that particular concept, within a body of information resources, can be indexed using this representative term.

Controlled vocabularies can apply to many different concepts, including subjects of resources, their formats, types, or the audiences for which the resources are intended.

Since many definitions for “controlled vocabulary” have been proposed or adopted in a variety of contexts and are frequently encountered in published resources on metadata and information management, it is useful to identify some significant issues addressed by this definition. First, for the purposes of this guide, controlled vocabulary values are always words or phrases ultimately derived from natural language in contrast to various types of alphanumeric strings or codes commonly associated with classification schemes. Values derived from a classification scheme usually require reference to the scheme outline itself for the meaning of such values to be apparent to human readers. Values derived from controlled vocabularies are immediately comprehensible to some degree by human readers.

The definition also recognizes that controlled vocabularies are used by at least two distinct types of users: first, by indexers, content analysts, or metadata creators as analytical or descriptive tools; and second, by information users or information managers for information retrieval or organization. Finally, the definition acknowledges that all controlled vocabularies have a predetermined, explicit, and coherent structure.

1.3 Why Are Controlled Vocabularies Useful?

Controlled vocabularies help indexers to describe information resources in a consistent manner, which fosters two outcomes.

- It allows users of those resources to find information efficiently.
- It allows information managers to separate unlike information resources and to bring together similar information resources.

The bringing together of similar resources is known as “collocation” in information science. One example of collocation occurs when information resources are grouped together under broad topic headings.

Although valuable for some purposes, full text searching is imprecise and often results in the retrieval of large amounts of unrelated information. In addition, creators of information resources do not always use the same terminology to identify concepts, topics, or subjects that users of resources in those areas use. For example, if a user were looking for studies or reports on the subject of heart attacks, he or she might perform a search using the term “heart attacks.” In a full text search of a large body of health-related Web resources, the search results would include any resources containing that phrase, regardless of the context in which the term is used. It might have been used in passing in an article about cancer or even as an interjection on a totally

unrelated topic. On the other hand, the search would not retrieve a study on acute myocardial infarctions if that study did not also contain the phrase “heart attacks” in the text.

For example, the use of a controlled subject vocabulary increases the consistency of search results and is a more efficient and precise way of searching than full text, as it retrieves all resources and *only* those resources that are about the subject being searched and have been indexed with the appropriate controlled vocabulary terms. This quality in search results is known as “precision.” When indexers apply subject terms to resources, they not only select the preferred term for a concept, but they also analyze the resources to determine what they are essentially about. Usually, this process results in a relatively small number of main concepts or subjects. The indexer then determines the preferred terms for these concepts. If all indexers have analyzed their resources correctly, searchers will retrieve precise or relevant results thanks to value-added, analytical human intervention.

By combining controlled vocabularies with sophisticated search algorithms, complex searches can be performed with various combinations of search terms.

It is important to note that controlled vocabularies are distinct and different from both the information technologies used to house them and the functions of those technologies. Furthermore, unless there is a technology designed to make appropriate use of the controlled vocabulary, it would be impossible to take advantage of any content indexing. For indexing to be functional from a user point of view, search engines must be configured to search for indexing terms specifically.

1.4 Why a common standard?

Controlled vocabularies can be as simple as short lists of designated words or phrases describing some aspect of a given information domain or as complex as a thesaurus with a very large number of preferred and non-preferred terms, including hierarchical and other semantic relations among terms. There is a wide variety of controlled vocabulary types within these two extremes. Some of these will be examined in detail in this guide.

However, it is important to note that a controlled vocabulary should only be as complex as is necessary to achieve its objective.

1.4.1 Authority Control, Classification Schemes, Controlled Vocabularies, and Taxonomies

A review of the relevant literature across communities of practice suggests that authority control, classification, controlled vocabularies, and taxonomies are, so far as structure and purpose are concerned, practically synonymous, if not actually identical in specific cases.

For the purposes of this guide, name authority files, subject authority files, and taxonomies are considered to be types of controlled vocabularies. Classification and classification schemes are not addressed in this guide because in most instances their characteristics and applications are so significantly different from controlled vocabularies as to require separate treatment. It is the intention of the Sub-Group to produce a separate document devoted to guidance on classification.

See also Appendix A: Glossary of Controlled Vocabulary Terminology.

1.4.2 Basic Controlled Vocabularies

The most basic controlled vocabulary is a list of designated words or phases relating to a particular information domain. These words or phases, called “indexing terms” or “descriptors,” may or may not require definitions or explanations, depending on their context or application.

However, even the most basic controlled vocabulary must follow rules for ensuring the consistent expression and formatting of terms. In effect, this is a crucial element of control in controlled vocabularies.

Small and basic controlled vocabularies may be adequately developed and maintained as a word processing document (i.e. a flat file).

An example of such a vocabulary is *Titles of Federal Organizations* (http://www.tbs-sct.gc.ca/pubs_pol/sipubs/tb_fip/titlesoffedorg_e.asp), a list intended to provide the definitive form of names for Canadian federal institutions. Another is the *Government of Canada Format Scheme* (http://www.tbs-sct.gc.ca/im-gi/mwg-gtm/fmt-fmt/docs/2003/schem_e.asp). This scheme includes terms that describe a wide variety of GC Web site formats for the representation of text, audio, video, and other media.

1.4.3 Controlled Vocabularies with Term Definitions

Slightly more complex controlled vocabularies may include the following:

- a word or phrase in parentheses after the first word or words in a term (also known as a parenthetical qualifier).

This qualifier is used to clarify the meaning of the term. In these cases, the entire text string, including the qualifier, constitutes the indexing term. In some vocabularies, square brackets are used instead of parentheses.

Example: Acquisitions (Businesses)

(Source: *Government of Canada Core Subject Thesaurus*— http://en.thesaurus.gc.ca/intro_e.html)

- term definitions explaining how to apply the term

Such definitions are usually prescriptive or restrictive, i.e. they either dictate or restrict the meaning of the term. The definition is not to be used as part of the indexing term.

Example: abstract A summary of a document or text

(Source: *Government of Canada Type Scheme*— http://www.tbs-sct.gc.ca/im-gi/mwg-gtm/typ-typ/docs/2003/schem/schem_e.asp)

- definitions interpreting the term, often known as scope notes, particularly in thesauri

Scope notes are indicative, explanatory, or prescriptive. Scope notes are usually indicated by the abbreviation "SN."

The purpose of a scope note is to:

- reduce ambiguity;
- increase clarity of meaning;
- indicate specificity;
- indicate restrictions on meaning;
- indicate the range of topics covered; and
- provide instructions to indexers.

The following is an example of a preferred term with a scope note:

Estates	SN	Investments, money, property or other valuables belonging to a deceased person. NOT to be used in the sense of: Landed property; individually owned piece of land containing a residence, esp. one that is large and maintained by great wealth.

(Source: *Government of Canada Core Subject Thesaurus*— http://en.thesaurus.gc.ca/intro_e.html)

1.4.4 Controlled Vocabularies with Preferred Terms

In natural language, a given concept is often represented or referred to by means of a wide range of words or phrases that express or imply a variety of contexts, shades of meaning, or application. Controlled vocabularies identify and restrict the many possibilities to a single or very limited meaning for the purposes of indexing, information retrieval, or information management. To ensure consistency in the description of concepts, some controlled vocabularies guide indexers and searchers from a set of possible terms representing a concept to the designated or preferred term for that concept. Consequently, all resources about, or pertinent to, that particular concept, within a body of information resources, can be indexed using this single representative term.

Non-preferred terms are also known as “lead-in” terms. Lead-in terms are synonyms of the preferred term that has been chosen from all terms as the only one authorized to represent the given concept. This is an essential element of “control” in controlled vocabularies.

An example of a preferred term, drawn from the *Government of Canada Core Subject Thesaurus* (http://en.thesaurus.gc.ca/intro_e.html), is:

Dangerous products	UF	Dangerous goods

UF is the abbreviation of “Used for.” This entry indicates that, in the context of the *Core Subject Thesaurus*, the concept of “Dangerous goods” should be described using the term “Dangerous products.” Those considering searching for “Dangerous goods” will be guided to the preferred term by finding the following entry:

Dangerous goods	USE	Dangerous products

1.4.5 Controlled Vocabularies With Hierarchies

Some controlled vocabularies provide links between terms representing broader and/or narrower concepts. Such hierarchies help users find terms that match their specific search needs. Controlled vocabularies that have purely hierarchical structures are frequently identified as taxonomies. Taxonomies that are designed for information management purposes may have multiple hierarchies, i.e. they may be polyhierarchical.

An example of a hierarchy is:

Dangerous products	BT	Products
	NT	Explosives

"BT" is the abbreviation of "Broader term" and "NT" is the abbreviation of "Narrower term." Hierarchical relationships are always reciprocal, e.g.:

Explosives	BT	Dangerous products
Products	NT	Dangerous products

1.4.6 Controlled Vocabularies With Related Terms

Related terms (abbreviated "RT") are terms with conceptual linkages to a given term. Related terms assist indexers and users to change or supplement their indexing or searching strategies respectively.

Example:

Dangerous products	RT	Chemicals

Related terms are always reciprocal, e.g.:

Chemicals	RT	Dangerous products

1.4.7 Complex Controlled Vocabularies

Complex controlled vocabularies typically include all of the features discussed in this section and may contain others as well. Those containing equivalents, hierarchical relationships, and associative relationships are generally designated as thesauri.

Examples of complex controlled vocabularies used within the GC are the

Government of Canada Core Subject Thesaurus (http://en.thesaurus.gc.ca/intro_e.html) and Health Canada's *E-Health Thesaurus* (http://www.hc-sc.gc.ca/ohih-bsi/res/thesaurus/thesaurus_alpha_e.html).

Complex controlled vocabularies normally require specialized software for their creation, operation, and maintenance.

1.5 Controlled Vocabularies for GC Metadata Elements

In the GC, the use of registered controlled vocabularies is required with the Dublin Core metadata elements shown in the table below. Schemes are lists of values from which metadata content may be selected.

Element Name	Element Mandatory or Optional	Sample Controlled Vocabularies
Subject	Mandatory	*GC Core Subject Thesaurus*
Audience	Optional	*GC Audience Scheme*
Coverage	Optional	*Canadian Geographical Names Database*
Format	Optional	*GC Format Scheme*
Type	Optional	*GC Type Scheme*

1.6 Optional Dublin Core elements

Library and Archives Canada (LAC) maintains a registry of authorized controlled vocabulary schemes. The function of this registry is twofold:

- To make standardized vocabularies available to search engines, information creators, and those involved in developing and maintaining vocabularies; and
- To provide a centralized mechanism for use in metadata elements for GC departments and agencies.

The criteria for registering controlled vocabularies include the following:

- Vocabularies must be created and maintained by trusted authorities.
- GC-owned vocabularies must be bilingual.
- GC-owned vocabularies must be publicly available on the World Wide Web.

What kinds of vocabularies are included?

- Controlled vocabularies, thesauri, or flat lists of terms may be registered.
- Vocabularies developed and maintained within the GC are registered.
- Well-known external vocabularies are also part of the registry.

Who may register a standardized vocabulary?

The departmental trusted authority or maintenance agency will submit a registration form to the standardized vocabulary registrar at LAC. LAC will then register well-known externally owned vocabularies on behalf of the GC.

How to register a vocabulary

Consult the supporting documentation, *Registering a Standardized Vocabulary*, on the LAC Web site at www.collectionscanada.ca/8/4/r4-293-e.html.

Send the completed form to the LAC Metadata Co-ordinator at meta_coord@lac-bac.gc.ca. You will receive e-mail confirmation when the registration has been completed.

For more information, see the registry's Web site at http://www.collectionscanada.ca/8/4/r4-293-e.html.

Chapter 2: Adopt, Buy, or Build?

When organizations acquire new information technologies, they are frequently faced with the choice of buying or building the required capabilities. The acquisition of controlled vocabularies as information management tools presents analogous choices. Specifically, organizations may:

a) Adopt and use an existing controlled vocabulary either as is or with alterations agreed to by the owner of the vocabulary;

b) Acquire and adapt a controlled vocabulary;

c) Extract term records in whole or in part from existing vocabularies as a basis for a new vocabulary or for developing another existing vocabulary; or

d) Build one from scratch.

Option a) includes the possibility of the vocabulary's owner being open to making adjustments to suit the needs of other organizations. Such adjustments could include added new, non-preferred, or relational terms or modifications to hierarchies.

Options b) and c) require the organization to take ownership of the controlled vocabulary insofar as development and maintenance are concerned. This implies considerably more investment of resources than would be required to use an existing controlled vocabulary. In fact, the creation of a new controlled vocabulary should not be considered until existing controlled vocabularies have been examined and assessed for potential use or modification. As will be seen in Chapter 4, creating even relatively small controlled vocabularies can be a very lengthy, complex, and expensive process. Further, the use of existing controlled vocabularies can increase the likelihood of metadata interoperability and reusability within organizational information infrastructures.

What are the factors that determine the choice to adopt, acquire/adapt, or build?

A question to consider is:

> Does at least one of the controlled vocabularies registered by LAC or the Dublin Core Metadata Initiative for use with the metadata elements employed by my department or agency supply the required terminology?

Terminology requirements are met if all concepts or subjects employed by, or of interest to, the organization's user community or contained in its information resources are represented in the controlled vocabulary, whether or not this representation is by means of preferred terms or non-preferred terms.

A number of controlled vocabularies have been developed and registered as a source of values for specific metadata elements. For example, the *Government of Canada Core Subject Thesaurus* is designated as the default thesaurus under TBITS 39.2. However, each department or agency must determine which controlled vocabulary is most suitable for its specific needs.

If none of the registered vocabularies can be used in its present form to describe its Web-based content in accordance with its needs, the department or agency should take the following steps, in this order:

a) Investigate other controlled vocabularies that may meet the organization's information management needs;

b) Investigate the possibility of requesting modifications to an existing vocabulary; and

c) Investigate the possibility of acquiring and making significant modifications to one or more existing vocabularies.

It is entirely possible that none of the registered vocabularies will meet these requirements. In that case, the organization should consider adapting an existing vocabulary, assuming that this is possible from a licensing or an intellectual property point of view. Alternatively, it will have to consider the possibility of committing resources to the development and maintenance of a new controlled vocabulary designed to meet its particular requirements.

The next two chapters will consider in some detail the two basic options of acquiring and adapting an existing controlled vocabulary or creating a new one.

Chapter 3: Assessing Existing Controlled Vocabularies

For assistance in carrying out the following analyses, seek advice from your departmental or agency library staff and from other information management and content experts in your organization.

3.1 Scoping the Problem

If an organization has identified an information management problem that could be solved by means of the application of a controlled vocabulary, then the following questions should be asked:

- Are there concepts that are not represented?
- Is there a problem with ambiguity or redundancy, e.g. are synonyms being used as indexing terms?
- Is there a lack of consistency in format or syntax, e.g. are abbreviations, acronyms, and expanded forms being used variously and indiscriminately together?
- Are lengthy descriptive phrases used instead of more concise and explicit terms?
- Are uncontrolled keywords employed?
- Are proper names consistently included or excluded?
- Are there problems with search results (lack of precision and recall; too many false drops)?
- Is there overall central editorial control?

If the answers to questions 1–5 are “yes” and those to questions 6–7 are “no,” then one may conclude that the indexing methodology employed has significant problems that will have serious negative impacts on information management and information retrieval within the organization.

If information resources are not currently being indexed using any controlled vocabulary or indexing methodology, then existing controlled vocabularies may be evaluated for potential use. The rest of this chapter is devoted to a methodology for evaluating such vocabularies.

3.2 Scoping the Solution

- Establish the scope of the required controlled vocabulary. Identify the principal terminology domains of the organization, including areas of concentration and of secondary importance.
- Determine the size of the information holdings to be indexed.
- Determine the scope of the information domain, i.e. broad and general or narrow and specific.
- Determine the rate of growth of the information holdings.

- Determine the rate of change or growth of language used in the information domain (e.g. the information domain may be subject to rapid expansion as the result of current research).
- Determine the number of queries directed at information holdings and the specificity of those queries.

In general, the greater the magnitude of each of the above factors, the larger and more complex the vocabulary is likely to be.

3.3 Selecting Available Controlled Vocabularies

As previously noted, a number of controlled vocabularies are registered on the LAC and Dublin Core Metadata Initiative Web sites. In addition, other controlled vocabularies are available on-line or in print.

For assistance in locating other potential controlled vocabularies, see the following Web sites:

- Michael Robert Middleton. Databases of Thesauri. Queensland University of Technology (http://sky.fit.qut.edu.au/~middletm/cont_voc.html#Databases)
- Willpower Information Management. Lists of thesauri. (http://www.willpower.demon.co.uk/thesbibl.htm#lists)
- Stephenson, Mary Sue. (2004). Indexing resources on the WWW. University of British Columbia. School of Library, Archival and Information Studies. [On-line bibliography]. Available at: http://www.slais.ubc.ca/resources/indexing/index.htm

Once located, controlled vocabularies should undergo a rigorous evaluation before they are selected for use. Initial considerations should include the following:

- Is the vocabulary available in both official languages?
- Is it publicly and freely available or is some form of licensing required?
- Are there any other restrictions on its use?
- Has it been well managed and is it likely to be maintained, as both language and terminological needs will evolve over time?
- Is it well documented and does it contain complete usage guidelines?
- Is its terminology generally appropriate in the GC and local departmental or agency contexts?
- Will vocabulary owners consider requests for the addition of new terms or the modification of existing ones?

To address the last point, the owner of a partially useful existing controlled vocabulary should be contacted with a view to exploring the possibility of making changes. Some negotiation may be

required, but the final enhanced vocabulary could be more useful and could have greater potential for reuse in the GC as a whole.

3.4 Evaluating a Controlled Vocabulary for Use

When as a result of the evaluation process a particular controlled vocabulary has been located and is a potential candidate for use, it may be evaluated using the following methodology:

- Select a statistically representative set of information resources held by the organization.
- Consider and ensure the appropriate representation of:
  - The size of the information holding;
  - The range of resource type;
  - The range of subjects, topics, and disciplines with which the organization is concerned; and
  - The group within the organization that creates or uses information resources.
- Index the representative set of documents with the target controlled vocabulary, using a predetermined indexing procedure.
- Assess the results using the following criteria:
  - Are there significant concepts that the target controlled vocabulary fails to represent?
  - Does the controlled vocabulary use language familiar to the user community of the organization?
  - Is the vocabulary capable of indexing to the required level of specificity?
  - Does the vocabulary include a sufficient number of lead-in terms (i.e. synonyms) that will guide the user community to the appropriate indexing terms, particularly if the vocabulary is large?
  - Taking into consideration all factors previously considered or assessed, will this controlled vocabulary serve the information management requirements of the organization?

It may be difficult or impossible to modify a partially useful controlled vocabulary, though it may be possible to acquire it in whole and then modify it independently of its parent organization. This would depend on several factors, including the intellectual property rights of the parent organization, costs, and technical considerations.

Since this option entails many of the decisions required in the maintenance of a vocabulary produced from the ground up, the guidelines presented in Chapter 4 should be reviewed and followed, should this approach be employed.

Chapter 4: Design and Construction of New Controlled Vocabularies

When no pre-existing controlled vocabulary can be adopted or adapted to meet ongoing information management requirements, then the creation of a new controlled vocabulary should be considered. The creation and maintenance of a controlled vocabulary, especially one containing a large number of terms, is a complex and resource-intensive undertaking requiring knowledgeable and experienced personnel and should not be undertaken lightly. Chapter 5 discusses the management of controlled vocabularies.

Readers seeking guidance on the nature and construction of large, complex controlled vocabularies should consult the following texts:

- Aitchison, J.A.; Gilchrist, A.; Bawden, D. (2000). Thesaurus construction and use: A practical manual (4th ed.). Chicago, IL: Fitzroy Dearborn.

  This textbook is widely acknowledged as the authority in professional indexing and controlled vocabulary communities. Although specifically directed at the design and construction of thesauri, many of its basic principles are applicable to other types of controlled vocabularies, including controlled vocabularies of simple structure and small size.

  This chapter has drawn extensively upon this publication, which will hereinafter be referred to as "AGB." Codes in parentheses refer to section numbers in that text.
- International Organization for Standardization (ISO). (1986). ISO 2788:1986 Documentation – *Guidelines for the Establishment and Development of Monolingual Thesauri*. 2nd ed. Geneva, Switzerland: International Organization for Standardization.
- International Organization for Standardization (ISO). (1985). ISO 5964:1985 Documentation – *Guidelines for the Establishment and Development of Multilingual Thesauri*. Geneva, Switzerland: International Organization for Standardization.
- National Information Standards Organization (NISO). (2003). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*. [ANSI/NISO Z39.19-2033]. Bethesda, MD: NISO Press. Available at: http://www.techstreet.com/cgi-bin/ detail?product_id=1171385

4.1 Definition of Scope and Purpose

Identify the nature and extent of the information domain. If the domain is large, identify areas of concentration and areas of secondary importance.

See also Section 3.2 for more information on determining specific needs and requirements for defining scope and purpose.

4.2 Determine Size and Structure

A key consideration is the size of the vocabulary. Size will determine structure (see Section 1.4). A vocabulary with a narrowly defined subject area and an extremely small number of terms (e.g. 10–30) that requires no synonyms and is unlikely to change over time may be relatively straightforward to develop and maintain.

However, if the body of knowledge to be addressed is large, varied, or expansive, the literary warrant approach should be considered. In general, literary warrant applies more to subject-oriented domains than to non-subject areas (e.g. characteristics such as formats or types of information).

The basic question is how specific the vocabulary should be to address the identified information needs. A vocabulary that is too general will retrieve too much irrelevant information; one that is too specific will retrieve too little.

The vocabulary should be sufficiently specific to address the information needs identified. “Granularity” is the term used to describe the level of detail reflected by the terms of a controlled vocabulary. A vocabulary of low granularity would represent many concepts under one term; one with high granularity would break out each concept into its own separate terms.

Generally, the broader the scope of the content to be described, the lower the granularity of the vocabulary. However, this may not be true if the vocabulary is very large. In addition, a highly focussed subject area often requires a high degree of granularity.

See also Section 1.4.

4.3 Construction Methodology (AGB / J4)

AGB identifies two approaches to selecting a construction methodology:

- Deductive

  > The examination of selected terms to find structure and to introduce vocabulary control is delayed until a sufficient number of terms have been collected. (AGB, p. 146)

- Inductive

  Terms are admitted to the [vocabulary] and are used in indexing as soon as they are encountered in the literature. Vocabulary control is applied at the outset and terms are allocated to one or more broader categories. Indexing may have to be revised [depending upon the functionalities of supporting information technologies]. (AGB, p. 146)

4.4 Scope of the Information Domain

It is essential to determine the resources to which the controlled vocabulary will be applied, as it is from this information domain that the terms (preferred and non-preferred) will be selected (for more information, see Section 3.2).

4.5 Literary Warrant and User Warrant (ANSI/NISO Z39.19-2003 5.3.5)

“Literary warrant” refers to a justification for selecting terms based on a significant frequency of occurrence of those terms in the information resources to be indexed. Literary warrant ensures that resources that could be indexed using the terms in the vocabulary actually exist. This ensures that unnecessary terms are not included in the vocabulary. It also ensures that the form of term selected as the preferred term will be the most commonly used in information resources.

“User warrant” refers to a justification for selecting terms based on words or phrases employed by users of information resources for information retrieval or information management. Evidence of such usage may be derived from search engine logs or interviews. User warrant ensures that the language of the vocabulary matches the language of the user community.

“Organizational warrant” refers to a justification for selecting terms based upon the business requirements and the business language used by an organization. Determining organizational warrant requires identifying the forms that are preferred by the organization that will use the controlled vocabulary. Organizational warrant ensures that the language of the vocabulary matches the needs and priorities of the organization.

The controlled vocabulary editorial policy should strike a balance among these three approaches. Literary warrant is usually the most reliable basis for term inclusion, but user warrant and organizational warrant may be good authorities for terminologies that will be easily recognized and understood by members of organizations or researchers. Both may be a good basis for establishing synonyms.

For more information on literary, user, and organizational warrant, see ANSI/NISO Z39.19-2003.

4.6 Sources of Terms (AGB / J5)

AGB note numerous sources of terminology, including:

- thesauri and lists of terms;
- classification schemes;
- encyclopedias, lexicons, dictionaries, and glossaries;
- terminological databanks; and
- manual and automated term extraction.

The following sources can be used specifically for the construction of GC-controlled vocabularies:

- GC publications in hard copy and on-line;
- GC classification schemes;
- Termium;
- user queries (search engine logs);
- user submissions; and
- editors' contributions.

4.7 Vocabulary Control

Vocabulary control is achieved in two ways:

a) Terms are deliberately restricted in scope to selected meanings.

b) When the same concept can be expressed by two or more synonyms, one of these terms is usually selected as the preferred term, which is then used consistently in indexing.

For more information, see AGB / J4, p. 146.

4.8 Concept Representation (AGB / D2)

It is useful to categorize indexing terms into generalized categories. AGB note that ISO 2788 identifies three main categories and their subdivisions.

- Concrete Entities
  - Things
  - Materials
- Abstract Concepts
  - Actions and events

  - Abstract entities
  - Properties of things, materials, and actions
  - Disciplines and sciences
  - Units of measurement
- Proper Nouns
  - Individual entities
  - Classes of one

(AGB / D2, p. 18).

4.9 Grammatical Forms (AGB / D3.1)

The editorial policy must establish the grammatical forms of words or phrases that are used as preferred and non-preferred terms in order to establish a uniform and consistent form for terms and their display.

- Nouns and noun phrases
  - Used for indexing terms; noun phrases are commonly used as modifiers
- Adjectives
  - Used only in compound indexing terms
- Adverbs
  - Used only in compound indexing terms
- Verbs
  - Usually excluded from subject vocabularies; probably required to some degree in functional vocabularies
- Initial articles
  - Initial articles should be avoided if possible

4.10 Singular and Plural Forms (AGB / D3.2 and ISO 5964)

It is strongly recommended that an editorial policy be established to ensure that terms appear with a high degree of consistency in either their singular or plural forms. Either form can be used in particular circumstances, but, by common convention, English terms generally appear in their plural forms. In contrast, French terms by convention are generally established in their singular form, but there are also many exceptions in French vocabularies that are determined by the meaning established for the term and usage. French terms by convention are generally

established in their masculine form where there is a possibility to have a masculine and a feminine form.

In addition, the following elements, in particular, should be considered.

- Concrete entities (AGB / D3.2.1 and ISO 5964 section 11.1.3)
  - In English, by convention, if something is countable (e.g. automobiles), use the plural form. In French, use the singular form (e.g. automobile).
  - For non-count nouns and collective nouns, use the singular form in English (e.g. sand) and French (e.g. sable).
- Abstract concepts (AGB / D3.2.2 and ISO 5964 section 11.1.3)
  - In English and French, by convention, use the singular form.
- Unique entities (AGB / D3.3.3 and ISO 5964 section 11.1.3)
  - In English and French, by convention, use the singular form.
- Co-existence of singulars and plurals (AGB / D3.2.4)
  - When the singular and plural forms have different meanings, a qualifying word or phrase is added in parentheses.
- Standardization (AGB / D3.2.5)
  - Standards for spelling, transliteration, romanization, etc. should be established in editorial policy, if applicable.

4.11 Punctuation and Capitalization (AGB / D3.5)

Editorial policy must establish a standardized approach to punctuation and capitalization practices.

- Parentheses
  - Parentheses are used only to enclose qualifiers.
- Hyphens
  - Avoid the use of hyphens wherever possible.
- Capitalization
  - Capitalization is used only for initial letters of proper names, by convention for the initial letters of indexing terms, and for acronyms.

4.12 Abbreviations, Initialisms, and Acronyms (AGB / D3.6)

AGB state that, “Abbreviations and acronyms may be used as indexing terms only when they have become so well known generally, or in a specialized field, that the abbreviated form is more familiar and the full form of the term is rarely used.” (p. 27)

For example, the term “AIDS” can be used as opposed to “Acquired Immune Deficiency Syndrome.”

4.13 Selection of Terms (AGB / D4)

Editorial policy must consider and decide whether or not the following will be included:

- loan words;
- neologisms, slang terms, and jargon;
- common names versus trade names;
- trade marks;
- popular names and scientific names;
- place names (the GC has registered geospatial vocabularies); and
- proper names of institutions, persons, etc. (identifiers).

4.14 Establishing the Meaning of Terms (AGB / D5)

See also ANSI/NISO Z39.19-2003, “Term Choice, Scope and Form.”

Natural language contains a multitude of words and phrases that designate particular concepts. These include homographs, homonyms, synonyms, quasi-synonyms, and lexical variants, among others, and all of them may cause problems of ambiguity and/or redundancy. By selecting a single term from among many possibilities, controlled vocabularies reduce these semantic problems to a minimum. Procedures for addressing these problems must be established as part of the editorial policy.

- **Scope Notes and Definitions (AGB / D5.2)**
  - Scope notes and definitions are textual devices for reducing or eliminating ambiguity and/or redundancy.

See Section 1.4.3 for more information.

4.15 Semantic Relationships in Controlled Vocabularies

It is important that terms are comprised of logical reciprocal relationships. The ANSI/NISO Z39.19-2003 standard emphasizes the importance of this: “Each relationship indicated between Term A and Term B must have a corresponding relationship from Term B to Term A. This rule must be observed for all types of relationships.” (43–8.11)

There are many types of relationships, as illustrated in AGB, that are necessary for the creation of controlled vocabularies. Some of these are outlined below.

- **Equivalence Relationships—Linguistic**

In the GC, all published or registered vocabularies must be available in both official languages. Additionally, all GC metadata records must have metadata in the language of the target information resource, under CLF 7.8. Therefore, in most cases, controlled vocabularies must have a field or data element for every preferred term that identifies the equivalent term in the other official language. This linguistic equivalent may not necessarily be a direct translation. Some terms in one language may have more than one equivalent in the other.

- **Equivalence Relationships—Synonymy, etc. (AGB / F1.1)**

Synonymous equivalence relationships occur when a semantic relationship is established between a preferred term and a non-preferred term so that the two terms, so far as indexing is concerned, refer to the same concept.

Types of synonymous relationships include:

- Popular names / scientific names;
- Common nouns / scientific names / trade names;
- Standard names / slang;
- Terms of different linguistic origin;
- Terms originating in different cultures sharing a common language;
- Competing names for emerging concepts; and
- Current / obsolete.

- **Equivalence Relationships as Lexical Variants**

- Variant spellings;
- Direct and indirect forms (e.g. “educational materials” or “materials, educational”);
- Abbreviations and full names.

- **Quasi-synonyms**

AGB define these as:

> ...terms whose meanings are generally regarded as different in ordinary usage, but they are treated as though they are synonyms for indexing purposes (e.g. “breweries” and “beverage industry”). Quasi-synonyms include terms having a significant overlap. The quasi-synonym device should be avoided as a means of reducing the size of the vocabulary by grouping together terms that ought to be treated as independent indexing terms, except in marginal subject areas.
> (AGB / F1.1.3, p. 52)

- **Upward (Generic) Posting**

…generic posting…treats narrower terms as if they are equivalent to, rather than as a species of, their broader terms. The effect is to reduce the size of the vocabulary, but at the same time to retain access via specific terms to the broader terms used to represent them.
(AGB / F1.1.3a, p. 53)

- **Hierarchical Relationships (AGB / F1.2; ANSI/NISO Z39.19-2003 8.3)**

If the controlled vocabulary is to include hierarchical relationships, it is highly recommended that the editorial policy consider and determine explicitly what type of hierarchical relationships will be established in the vocabulary.

Hierarchical relationships are based on levels of superordination and subordination, where the superordinate term represents a class or a whole and is labelled as the broader term (BT), and subordinate terms refer to its parts, or narrower aspects of the class (NT).

The classic rule for validity in hierarchical relationships may be stated as: “Terms are hierarchically related only if both are members of the same fundamental category (facet); that is, they represent *entities*, *activities*, *agents*, or *properties*, etc.” However, in most instances of the “subjective” hierarchies described below, this rule is violated.

Hierarchical relationships include:

- The generic relationship, which “identifies the link between a class and its members or species” (ANSI/NISO Z39.19-2003 8.3.1; p 47);
- The hierarchical whole-part relationship, in which “one concept is inherently included in another, regardless of context, so the terms can be organized into logical hierarchies, with the whole treated as a broader term (ANSI/NISO Z39.19-2003 8.3.3., p. 49);

- The instance relationship, which "identifies the link between a general category of things or events, expressed by a common noun, and an individual instance of that category" (ANSI/NISO Z39.19-2003 8.3.2, p. 48); and
- The polyhierarchical relationship, in which "concepts belong, on logical grounds, to more than one category." (ANSI/NISO Z39.19-2003 8.3.4., p. 50).

Alternatively, Professor Michèle Hudon[1] has classified hierarchies as follows:

- Generic ("is a …");
- Partitive (partonomy) ("is part of …");
- Contextual / conventional ("is found in…context"); and
- Instance ("is an example of…").

In this arrangement, *polyhierarchies* are not included because they are considered to be simply more complex instances of either *generic* or *partitive* relationships.

Hudon further characterizes these types by distinguishing *generic* and *partitive* relationships as "objective" (i.e. rule-driven) and the *contextual* and *instance* types as "subjective" (i.e. established by particular circumstances, contexts, or business requirements). The principal disadvantage of the "subjective" approach is that the underlying logic of the relationship is not easily determinable by users (human or machine).

Hudon also characterizes the "power of hierarchy" as having the following characteristics:

- Complete and comprehensive information;
- Inheritance of attributes;
- Inference—the hierarchy allows reasoning from incomplete evidence;
- Definition—the hierarchical structure provides a way of expressing how an entity is like others and how it is different from others; and
- High-level view and holistic perspective: the hierarchical structure provides a visualization of a domain or phenomenon.

1. Hudon, Michèle. (2004). "What do we mean 'Taxonomy'?" In: "Public Sector Taxonomy Day." Hosted by the Council of Federal Libraries, March 9, 2004, Congress Centre, Ottawa ON.

As stated above, the validity of the "objective" hierarchical relationships may be established by the application of logical rules. For *generic* relationships, the rules may be stated as follows:

> The *generic* relationship must have the mathematical property of inheritance, whereby what is true of a given class is also true of all the classes subsumed under it. The relationship is correct if both the genus and species are of the same fundamental category (facet). The generic relationship applies to types of actions, properties, and agents, as well as to types of things (entities).

- The logical test:
  A is a type of B; A is always a type of B.
  Example:
- The all/some test:
  Some members of a class X are entities Y, while all entities Y are members of class X.
  Example:
  - Some rodents are squirrels; all squirrels are rodents. (valid)
  - Some pests are squirrels; all squirrels are pests. (invalid)

The *whole-part* (partitive) hierarchical relationship has very precise limitations established by ISO and NISO standards. In general, the relationship is valid if "the name of the part implies the name of the possessing whole in any context." Four types have been identified, as follows:

- Systems and organs of the body;
- Geographical location;
- Discipline or field of study; and
- Hierarchical social structure.

The logical test:

- A is an element, subset, aspect, or object of B or vice versa and at least one of the following is true:
  - If A exists, then B exists.
  - If B exists, then A exists.

In all other circumstances, the whole-part relationship will be an associative one, i.e. the terms will be related terms (RTs).

- **Associative Relationships (AGB / F1.3)**

As the phrase suggests, associative relationships (related terms, i.e. RTs), have a close or significant semantic relationship but one that is neither hierarchical nor equivalent (synonymous). An associative relationship provides a suggestion to an indexer or searcher to consider terms that are commonly linked in various ways in information resources, fields of knowledge, or in natural language. Two general rules are:

  - One of the terms should be strongly implied, according to the frames of reference shared by the users of the thesaurus, whenever the other is employed as an indexing term; and
  - One of the terms is a necessary component in any definition or explanation of the other.

- **Types of Associative Relationships**
  - Same category
    - Terms with overlapping meanings (e.g. Ships and Boats)
    - One concept is derived from the other
  - Different categories
    - The whole-part associative relationship (e.g. Harbours—Wharfs)
    - A discipline or field of study versus the objects or phenomena studied (e.g. Ornithology—Birds)
    - An operation or process versus the agent or instrument (e.g. Photocopying—Photocopier)
    - An occupation versus the person in the occupation (e.g. Nursing—Nurse)
    - An action versus the product of the action (e.g. Photocopying—Photocopies)
    - An action versus its patient (e.g. Food inspection—Food)
    - Concepts versus properties of those concepts (e.g. Paint—Colour)
    - Concepts versus the origins of those concepts (e.g. Children—Parents)
    - Concepts versus causal dependence (e.g. Explosives—Explosions)
    - A thing or action versus its counter-agent (e.g. Head injuries—Helmets)
    - Raw material versus product (e.g. Iron ore—Steel)
    - An action versus an associated property (e.g. Food inspection—Food safety)
    - A concept versus its opposite (antonym not treated as a quasi-synonym) (e.g. Imports—Exports)

Compound Terms

In general, controlled vocabulary terms should be single words representing single concepts. This goal is not always attainable, especially in larger vocabularies designed to index complex information domains. If compound terms (word phrases) must be employed as indexing terms, then it is highly recommended that editorial policy establish rules or guidelines for the formatting or admission of such terms into the vocabulary. The *Art and Architecture Thesaurus* has established guidelines for compound terms that are a good starting point for establishing an editorial standard in this regard. A summary of these guidelines is available as an appendix in AGB.

4.16 Publication and Implementation

If commercial thesaurus construction software has been used as an editing tool, then some degree of quality assurance will have been carried out automatically. Spell checking and other text editing procedures are nevertheless recommended. Testing is also recommended and may be carried out using the methodology described in Section 3.4. A final preparation for publication should include a written introduction that contains the following elements from ANSI/NISO Z39.19-2003:

- Statement of purpose;
- Statement of subject coverage;
- Identification of ongoing editorial authority;
- Number of indexing terms;
- Vocabulary control standards;
- Structure and interrelationships;
- Thesaurus layout and display;
- Abbreviations, punctuation;
- Operational use; and
- Updating and maintenance.

Chapter 5: Management of Controlled Vocabularies

5.1 Documentation and Communication

A fundamental requirement for the use of controlled vocabularies in metadata is that ownership and accountability for maintenance and sustainability is clearly documented. This will include instructions on access and procedures for using external vocabularies or explicit policy on the development and maintenance of internal vocabularies supported by clear governance structures.

All vocabulary policies and procedures should be thoroughly documented and continuously updated. Major changes in vocabulary policies or procedures that will affect any users should be communicated to them as soon as possible.

5.2 Governance and Policy Issues

The governance structure for the controlled vocabulary should be established at the outset. This structure includes the following:

- Ownership: In what individual or group is ownership and control vested? (Ownership should be communicated clearly to all using the vocabulary.)
- Editorial policy: How are terms included, modified, etc.?
- Indexing policy: How is the vocabulary to be applied to information resources?
- Maintenance policy: How will the vocabulary be managed over time?
- Publication of the vocabulary and its supporting documentation.

5.3 Resource Issues

Resource issues that must be considered include:

- Assessing the development time and costs of creation and maintenance;
- Creating a business case; and
- Obtaining the necessary resources (see also Section 5.6).

5.4 Technology

Technology issues include:

- Editing and maintenance tools;
- Applications for Web publishing;
- Web services applications; and
- Automated metadata tools.

5.5 Training

Depending upon the complexity of the controlled vocabulary and of the information technology that supports it, training packages may be required for indexers, searchers, or information managers.

5.6 Maintenance

Even the simplest controlled vocabulary will need to be maintained as the information domain changes or expands. Consequent changes in terminology can be labour-intensive and time-consuming. Therefore they must be identified as part of ongoing resource requirements.

If the vocabulary is publicly accessible, it is open to audit, and liability issues could arise if it is not properly maintained.

An official maintenance policy should be established addressing such aspects as:

- How changes to the information requirements will be monitored—these may include changes such as new legislation, modifications to the objectives of the department or agency, or new users and/or purposes for the vocabulary;
- How changes within the information domain will be assessed over time;
- How terms are added or modified (specific processes must be identified);
- How changes to terms are tracked over time;
- How changes to local technology will affect operation; and
- Whom to inform when changes are made (the vocabulary may be used by many different users and organizations that need to be informed of modifications).

5.7 Professional Assistance

The development and maintenance of controlled vocabularies requires highly specialized skill sets. While graduates of Library and Information Science programs and experienced librarians are likely to be familiar with the concepts involved, it cannot be assumed that all of them will have the requisite expertise.

The necessary expertise can be obtained through a combination of:

- Specialized courses (e.g. technical writing) in which the theory and its practical application in areas such as indexing are addressed in depth; and
- Experience in the construction and maintenance of actual controlled vocabularies, particularly if the planned vocabulary will be large and complex.

Other considerations include:

- Knowledge of the application / business environment;
- Advanced computer literacy;
- Experience in the areas of information system design, organization, and usability;
- Knowledge of and the ability to apply international standards; and
- Knowledge of the software package(s) being used.

External contracting can be an appropriate option at the creation stage. Appropriate personnel can be hired, trained, or contracted to maintain the vocabulary once constructed.

5.8 Tools

For the most basic controlled vocabularies, such as flat files with few if any synonyms or relationships between terms, common spreadsheet or word processing packages may suffice. However, for more complex requirements, more specialized software should be investigated. Some commercial packages are available.

In reviewing potential tools, considerations include the following:

- Ensuring the tools will co-exist with the existing and emerging technological environment within the department or agency.
- If the controlled vocabulary is registered and publicly available, the tool must support such access requirements.
- Tools with English and French interfaces are preferred.
- If possible, tools compliant with ISO standards should be selected.

The Dublin Core Metadata Initiative maintains a Web site devoted to tools and software (http://dublincore.org/tools). See also *Software for Building and Editing Thesauri* at http://www.willpower.demon.co.uk/thessoft.htm.

Appendix A: Glossary of Controlled Vocabulary Terminology

AGB	Aitchison, J.A.; Gilchrist, A.; Bawden, D. (2000). Thesaurus construction and use: A practical manual (4th ed.). Chicago, IL., Fitzroy Dearborn.
Associative term	See “Related term.”
Boolean search	A search in which terms can be combined with Boolean operators such as “AND,” “OR,” or “BUT NOT”.
Broader term	A term to which another term or multiple terms are semantically subordinate in a hierarchy.
Classification scheme	Systematic identification and arrangement of business activities and/or records into categories according to logically structured conventions, methods, and procedural rules represented in a classification system (Ref.: ISO 15489-1 Information and Documentation—Records Management—Part 1: General).
Controlled vocabulary	A list of standardized terminology, words, or phrases used for indexing or content analysis and information retrieval usually in a defined information domain. It is characterized by consistent format and syntax and may include synonyms and cross-references. In a controlled vocabulary, one of a set of possible terms representing a concept can be used as the representative term for that concept. Consequently, all resources about, or pertinent to, that particular concept, within a body of information resources, can be indexed using this representative term.
Exhaustivity	This designates the range of concept coverage of terms in a controlled vocabulary. If the terms cover all of the concepts included in the information domain, then the controlled vocabulary is exhaustive.
Facet	A clearly defined, mutually exclusive, and collectively exhaustive aspect, property, or characteristic of a class or specific subject.
False drop	An item in the search results, considered as being out of context.
Flat file	Simple lists of terminology without synonyms or non-preferred terms.
Format	The technology in which a resource is encoded.

Functional vocabulary	A controlled vocabulary that describes the functions and operations of an organization.
Granularity	This refers to the level of specificity with which content is described. A vocabulary of low granularity represents many concepts under one term; one with higher granularity breaks out the concepts into their own separate terms.
Hierarchies	A system of ranked terms in which a superordinate or higher term is broader in semantic scope than a subordinate or lower term.
Homographs (polysemes)	Words having the same spelling as another but differing in origin and meaning.
Homonyms	Words that can denote different meanings or words having the same sound but different meanings.
Indexing	An operation intended to represent the results of the analysis of a document by means of a controlled or natural indexing language.
Information domain	A well-defined area of knowledge, including the information resources pertaining thereto.
Keyword	A word occurring in the natural language of a document that is considered significant for indexing and retrieval.
Lead-in Term	See “Non-preferred terms.”
Literary warrant	This is a justification for selecting terms based on a significant frequency of occurrence of those terms in the information resources to be indexed. Literary warrant assures that resources that could be indexed using the terms in the vocabulary actually exist. This tends to ensure that unnecessary terms are not included in the vocabulary. It also ensures that the form of term selected as the preferred term will be the most common term used in information resources.
Mapping	The process of relating the terms in one controlled vocabulary to those in another.
Metadata	Structured information about the contents and/or nature of information resources.

Name authority file	A list or file that is maintained to ensure the consistency of indexing and that establishes the authoritative form of a corporate, geographic or personal name that is to be used to index records. A name authority file may contain variant forms of names that are cross-referenced to the authoritative form of a name.
Narrower term	A term that is subordinate to another term in a hierarchy.
Non-preferred term	This designates a synonym, quasi-synonym, lexical variant, or equivalent term to a preferred term; non-preferred terms do not represent concepts in a controlled vocabulary, but instead guide the user to preferred terms.
Post-co-ordination	The combination of simple terms to represent more complex concepts at the time of search or of indexing.
Precision	A ratio that measures the success of a search (No. of relevant items returned / Total no. of items returned = x) Precision has an inverse relationship to recall. Precision can be increased by increasing the specificity of index terms.
Pre-co-ordination	The combination of simple terms in compound phrases or terms in a controlled vocabulary to represent complex concepts in advance of indexing or of searching.
Preferred term	A term used consistently when indexing to represent a given concept; also known as a descriptor.
Quasi-synonym	A term whose meanings are generally regarded as different in ordinary usage, but that are treated as though they were synonyms for indexing purposes.
Recall	A ratio that measures the success of a search. Recall has an inverse relationship to precision. (No. of relevant items returned / Total no. of relevant items in the collection = x)
Related term	A term that has semantic associations with another in the context of a controlled vocabulary.
Scope note	This refers to an explanation, clarification, definition, or usage limitation for a term in a controlled vocabulary; a scope note is not part of the term so far as indexing is concerned.
Specificity	The exactness with which a term covers a concept; increasing specificity increases precision but may decrease recall.

Synonym	A term having a different form but exactly or very nearly the same meaning as another term.
Syntax	The form in which the terms of a controlled vocabulary are represented.
Taxonomy	A hierarchical arrangement of topics that imposes topical structure on information in a specific body of knowledge.
Term	A word or phrase used in a controlled vocabulary.
Thesaurus	This refers to a controlled vocabulary arranged in a known order (not necessarily alphabetic) in which equivalence (Use; Use for), homonymous (variant spellings), hierarchical (Broader Term; Narrower Term), and associative (Related Term) relationships among terms are clearly displayed and identified by standardized relationship indicators. A thesaurus also contains synonyms or “lead-in” terms that may be used as the conceptual point of entry by searchers or indexers, instead of the designated indexing terms.
Top term	A term that has no broader term, i.e one that occupies the “top” of its semantic hierarchy.
User warrant	This term refers to a justification for selecting terms based on words or phrases employed by users of information resources for information retrieval or information management. Evidence of such usage may be derived from search engine logs.

Appendix B: Selected Bibliography of Controlled Vocabulary Resources

Adams, Katherine C. (2000). Immersed in structure: The Meaning and Function of Taxonomies. In: Internetworking 3:2, August 2000. Available at: http://www.internettg.org/newsletter/aug00/article_structure.html.

Aitchison, Jean. (1970). The thesaurofacet: A multipurpose retrieval language tool. In: Journal of Documentation. 26:1, September 1970. 187–203.

Aitchison, J.A.; Gilchrist, A.; Bawden, D. (2000). Thesaurus construction and use: A practical manual (4th ed.). Chicago, IL., Fitzroy Dearborn.

Alani H.; Jones, C.; Tudhope, D. (2000). Associative and spatial relationships in thesaurus-based retrieval. Berlin: Springer. Lecture Notes in Computer Science. In: Proceedings (ECDL 2000) 4th European Conference on Research and Advanced Technology for Digital Libraries, (J. Borbinha, T. Baker eds).

Australia. New South Wales. State Recordkeeping. (2000). Compiling a functional thesaurus to merge with Keyword AAA. Available at: http://www.records.nsw.gov.au/publicsector/rk/rib/mergeaaa.htm.

Batty, David. (1989). Thesaurus construction and maintenance: A survival kit. In: Database. 12:1. 13–20.

Batty, David. (1998). WWW – Wealth, Weariness or Waste: Controlled vocabulary and thesauri in support of online information access. In: D-Lib Magazine, November, 1998. Available at: http://www.dlib.org/dlib/november98/11batty.html.

Bedford, Denise A. D. (2003). Programmatic Approaches to Metadata Capture & Controlled Vocabulary Creation. Presentation to 20th Anniversary Workshop British Library, London England September 29th, 2003. Available at: http://www.multites.com/presentations/ControlledVocabularyCreation.ppt.

Bedford, Denise A. D. (2003). Use of Thesauri, Classification Schemes & Rule-based Engines in the World Bank Catalog. Presentation to 20th Anniversary Workshop British Library, London England September 30th, 2003. Available at: http://www.multites.com/presentations/WorldBank.ppt.

Blocks D.; Binding C.; Cunliffe C.; Tudhope D. (2002). Qualitative evaluation of thesaurus-based retrieval. Berlin: Springer. Lecture Notes in Computer Science. In: Proceedings (ECDL 2002) 6th European Conference on Research and Advanced Technology for Digital Libraries, (M. Agosti, C. Thanos eds.), 346–361. Available at: http://www.glam.ac.uk/socschool/research/hypermedia/publications/presentationdocs/ecdl.pdf.

Broeder, Dann; Offenga, Freddy; Willems, Don. (2002). Metadata Tools Supporting Controlled Vocabulary Services. Available at: http://www.mpi.nl/IMDI/documents/2002 LREC/Metadata Tools Supporting Controlled Vocabulary Services.pdf.

Brown, Fred. (1998). Vocabulary Links: //Thesaurus Design for Information Systems – Seminar by Dr. Bella Hass Weinberg. Available at: http://www.allegrotechindexing.com/article02.htm.

Buckland, Michael. (1999). Vocabulary as a central concept in Library and Information Science. Preprint of paper published as “Vocabulary as a Central Concept in Library and Information Science” In: Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3, Dubrovnik, Croatia, 23–26 May 1999. Ed. by T. Arpanac et al. Zagreb: Lokve, pp 3–12. ISBN 953-6003-37-6. Available at: http://www.sims.berkeley.edu/~buckland/colisvoc.htm.

Chan, Lois Mai; Zeng, Marcia Lei. (2002). Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: A Methodological Analysis. IFLA. In: 68th IFLA Council and General Conference August 18–24, 2002. Available at: http://www.ifla.org/IV/ifla68/papers/008-122e.pdf.

Clack, Doris Hargrett. (1990). Authority control: Principles, applications and instructions. Chicago: American Library Association.

Craven, T. (2001). Thesaurus Construction. London, Ont., University of Western Ontario. Available at: http://instruct.uwo.ca/gplis/677/thesaur/main00.htm.

Cullinan Sievert, MaryEllen; Patrick, Timothy B.; Reid, John C. (2001). Need a bloody nose be a nosebleed? or, lexical variants cause surprising results. In: Bulletin of the Medical Library Association. 89:1, 2001. 68–71. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=31706, http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=31706&action=stream&blobtype=pdf.

Doerr, Martin. (2001). Semantic Problems of Thesaurus Mapping. In: Journal of Digital Information 1:8. Available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/.

Fast, Karl; Leise, Fred; Steckel, Mike. (2002–2004). What is a controlled vocabulary? Boxes and Arrows. Available at: http://www.boxesandarrows.com/archives/what_is_a_controlled_vocabulary.php.

Ganzmann, Jochen. (1990). Check-list for thesaurus software. In: International classification. 17:3/4. 155–157. Available at: http://www.willpower.demon.co.uk/criteria.htm.

Ganzmann, Jochen. (1990). Criteria for the evaluation of thesaurus software. In: International Classification. 17:3/4. 148–157. Available at: http://www.willpower.demon.co.uk/ganzmann.htm.

Gilchrist, Alan. (2001). Mapping Terminologies – Problems. University of Strathclyde. HILT Project. HILT Workshop, Glasgow, June 19, 2001. Available at: http://hilt.cdlr.strath.ac.uk/Dissemination/Presentations/Alan Gilchrist 2000.ppt.

Gilchrist, A. D. (1994). Classification and thesauri. London, England. ASLIB. In: Fifty Years of Information Progress: a Journal of Documentation Review. Vickery, B. (ed.). 85–118.

Hagedorn, Kat. (2001). Extracting value from automated classification tools: The role of manual involvement and controlled vocabularies. Argus Associates, Inc. Available at: http://argus-acia.com/white_papers/classification.pdf.

Hannon, Kevin. (2004). One large or many small? Executing taxonomies for large organizations. Collaboration Expedition Workshop #31: Joint Workshop on Multiple Taxonomies, Arlington, VA, April 28, 2004. Available at: http://colab.cim3.net/file/work/Expedition_Workshop/2004-04-28_Multiple_Taxonomies/Hannon_20040428.ppt.

Haynes, David. (2004). Metadata for information management and retrieval. London: Facet Publishing.

Hudon, Michèle. (1998). Compatibility and identity are not synonyms: Conceptual structures in multilingual thesauri. In: Knowledge Organization. 25:4. 152–155.

Hudon, Michèle. (1997). Multilingual thesaurus construction: Integrating the views of different cultures in one gateway to knowledge and concepts. In: Knowledge organization. 24:2. 84–91.

Hudon, Michèle. (2001). Relationships in multilingual thesauri. In: Relationships in the Organization of Knowledge. 67–80.

Hudon, Michèle. (2004). “What do we mean…‘Taxonomy’?” In: “Public Sector Taxonomy Day.” Hosted by the Council of Federal Libraries, March 9, 2004, Congress Centre, Ottawa ON.

Hunter, Jane. (2001). MetaNet – A Metadata Term Thesaurus to Enable Semantic Interoperability between Metadata Domains. In: Journal of Digital information. 1(8). Available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/.

International Organization for Standardization (ISO). (1986). ISO 2788:1986 Documentation – Guidelines for the establishment and development of monolingual thesauri. 2nd ed. Geneva, Switzerland: International Organization for Standardization.

International Organization for Standardization (ISO). (1985). ISO 5964:1985 Documentation – Guidelines for the establishment and development of multilingual thesauri. Geneva, Switzerland: International Organization for Standardization.

Jonghoon, Lee; Dubin, David. (2002). Vocabulary mapping in the NASA ADS: Prospects for practical subject access. Prague, Czech Republic. In: Library and Information Services in Astronomy IV July 2–5, 2002. Available at: http://www.eso.org/gen-fac/libraries/lisa4/Dubin1.pdf

Kim, Youngin. (1998). Sensitivity of Entry Vocabulary Modules to Subdomains. University of California, Berkeley. School of Information Management and Systems. Metadata Research Program. Available at: http://metadata.sims.berkeley.edu/subdomain.html.

Kim, Youngin; Norgard, Barbara. (1998). Adding Natural Language Processing Techniques to the Entry Vocabulary Module Building Process. University of California. School of Information Management and Systems. Metadata Research Program. Available at: http://metadata.sims.berkeley.edu/nlptech.html.

Koch, Traugott. (1996; 2001). Controlled vocabularies, thesauri, classification schemes. [Online bibliography]. Available at: http://www.lub.lu.se/metadata/subjects-help.html.

Koch, Traugott. (2000). Quality-controlled subject gateways: definitions, typologies, empirical overview. In: Online Information Review. 24:1. 24–34. [Online bibliography]. Available at: http://www.lub.lu.se/tk/demos/SGin.html.

Kramer, Ralf; Nikolai, Ralf; Habeck, Corinna. (1997). Thesaurus federations: Loosely integrated thesauri for document retrieval in networks based on Internet technologies. In: International Journal of Digital Libraries. 1:1997. 122–131.

Language Independent Metadata Browsing of European Resources (LIMBER). (2001). [LIMBER thesaurus]. Language Independent Metadata Browsing of European Resources (LIMBER). [This thesaurus description report describes the methods used to produce the monolingual version of the European Language Social Science Thesaurus (ELSST). It also contains the actual listings of the 46 main hierarchies and an alphabetic listing of all the 1,523 terms within those hierarchies, the included top terms of other hierarchies and stand-alone terms]. Available at: http://www.limber.rl.ac.uk/Internal/Deliverables/D4.1 Thesaurus/D4.1_V2_final.doc.

Maple, Amanda. (1995). Faceted Access: A Review of the Literature. Available at: http://www.music.indiana.edu/tech_s/mla/facacc.rev; and http://www.musiclibraryassoc.org/BCC/BCC-Historical/BCC95/95WGFAM2.html.

Matthews, Brian. (2003). Migrating Thesauri to the Semantic Web. W3C Semantic Tour, London Zoo. Presentation in HTML. Available at: http://www.w3c.rl.ac.uk/pastevents/matthews/semantictour/title.html.

Miller, Ken. (2000). A way with words: Thesauri assisted searching. Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. Work session on statistical metadata. (Washington, 28–30 November 2000). (Working Paper No. 7). Available at: http://www.unece.org/stats/documents/2000.11.metis.htm.

Miller, Paul. (2000). I say what I mean, but do I mean what I say? In: Ariadne. 23, Mar. 22. Available at: http://www.ariadne.ac.uk/issue23/metadata/ - 26.

Milstead, Jessica L. (2000). About thesauri. Available at: http://www.bayside-indexing.com/Milstead/about.htm.

Milstead, Jessica. (1998). NISO Z39.19: Standard for Structure and Organization of Information Retrieval Thesauri. Kensington, CA: Bayside Indexing Service. Paper presented at the Taxonomic Authority Files Workshop, Washington, DC, June 23, 1998. Available at: http://www.bayside-indexing.com/Milstead/z39.htm.

National Archives of Australia. (2003). Developing a Functions Thesaurus. Available at: http://www.naa.gov.au/recordkeeping/control/functions_thesaur/thesaurus.pdf.

National Archives of Australia. (2000). Assessment of the Keyword AAA Thesaurus. Available at: http://www.naa.gov.au/recordkeeping/control/KeyAAA/aaa/assessment.html.

National Information Standards Organization (NISO). (2003). Guidelines for the Construction, Format, and Management of Monolingual Thesauri. [ANSI/NISO Z39.19-2033]. Bethesda, MD: NISO Press. Available at: http://www.techstreet.com/cgi-bin/detail?product_id=1171385.

Omelayenko, Borys. (2002). Integrating Vocabularies: Discovering and Representing Vocabulary Maps. Talk at the 1st International Semantic Web Conference Sardinia, Italy, 10–12 June, 200. Available at: http://citeseer.ist.psu.edu/699976.html.

O'Neill, Edward T.; Chan, Lois Mai. (2003). FAST (Faceted Application of Subject Terminology): A simplified vocabulary based on the Library of Congress Subject headings. In: IFLA Journal. 29:4. 336–342. Available at: http://www.ifla.org/V/iflaj/ij-4-2003.pdf.

Pidcock, Woody. (2003). What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? Metamodel.com. Available at: http://www.metamodel.com/article.php?story=20030115211223271&mode=print.

Québec. Conseil du trésor. Sous-secrétariat à l'inforoute gouvernementale et aux ressources informationnelles. (1998) Thésaurus de l'activité gouvernementale : Orientations et principes directeurs. Québec. Conseil du trésor. Sous-secrétariat à l'inforoute gouvernementale et aux ressources informationnelles. Collection en ingénierie documentaire 10. Adresse : http://www.services.gouv.qc.ca/fr/publications/enligne/administration/ingenierie/thesaurus_orientations.pdf.

Robinson, Catherine; Knight, Janet. (2000). Contemporary Recordkeeping: The Records Management Thesaurus – Response. Australia. New South Wales. State Recordkeeping. Article in response to paper delivered at the 1997 Records Management Association of Australia's National Convention. Available at: http://www.records.nsw.gov.au/publicsector/rk/aaa/response.htm.

Saadani, Lalthoum; Bertrand-Gastaldy, Suzanne. (2000). Cartes conceptuelles et thésaurus : Essai de comparaison entre deux modèles de représentation issus de différentes traditions disciplinaires. In: CAIS 2000: Dimensions Of A Global Information Science: Canadian Association for Information Science, Proceedings of the 28th Annual Conference. Available at: http://www.slis.ualberta.ca/cais2000/saadani.htm.

Slavic, Aida. (2000). A Definition of Thesauri and Classification as Indexing Tools. Dublin Core Metadata Initiative. Available at: http://dublincore.org/documents/2000/11/28/thesauri-definition/.

Stephenson, Mary Sue. (2004). Indexing resources on the WWW. University of British Columbia. School of Library, Archival and Information Studies. [On-line bibliography]. Available at: http://www.slais.ubc.ca/resources/indexing/index.htm.

Willpower Information. (1992). Thesaurus principles and practice. Available at: http://www.willpower.demon.co.uk/thesprin.htm.

Vizine-Goetz, Diane et al. (2004). Vocabulary Mapping for Terminology Services. Southampton, UK: In: Journal of Digital Information. Volume 4 Issue 4. Available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/.

Warner, Amy J. (2004). A Taxonomy Primer. Available at: http://www.lexonomy.com/publications/aTaxonomyPrimer.html.

Will, Leonard. Costs of vocabulary mapping. Available at: http://hilt.cdlr.strath.ac.uk/Dissemination/Presentations/Leonard Will.ppt.