HANDBOOK OF TERMINOLOGY MANAGEMENT

HANDBOOK OF TERMINOLOGY MANAGEMENT

VOLUME 2

Application-Oriented Terminology Management

Compiled by

SUE ELLEN WRIGHT
Kent State University

GERHARD BUDIN
University of Vienna

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM / PHILADELPHIA

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Handbook of terminology management / compiled by Sue Ellen Wright, Gerhard Budin.
p. cm.
Contents: v. 2. Application-Oriented terminology management.
1. Technology--Terminology--Handbooks, manuals, etc. I. Wright, Sue Ellen, 1943-. II. Budin, Gerhard.
T11.H27 1997
401'.4--dc21 01-7100
ISBN 90 272 2155 3 (Eur.) / 1 55619 509 5 (US) (Vol. 2; alk. paper)
ISBN 90 272 2149 9 (Eur.) / 1 55619 502 8 (US) (Set; alk. paper) CIP

John Benjamins Publishing Company • P.O.Box 75577 • 1070 AN Amsterdam • The Netherlands
John Benjamins North America • P.O.Box 27519 • Philadelphia PA 19118-0519 • USA

CONTENTS

VOLUME II: APPLICATIONS-ORIENTED TERMINOLOGY MANAGEMENT

INFORMATION BOXES

APPENDICES

List of Illustrations

Chapter 6.1, Computer-assisted Thesaurus Management

Chapter 6.2, Terminology and Indexing

Chapter 6.3, Terminology and Information Retrieval

Chapter 7.1, Terminology in Technical Writing

For the convenience of readers who may not have Volume I of the Handbook of Terminology Management readily at hand, the Table of Contents from that volume is reproduced below.

VOLUME I: BASIC ASPECTS OF TERMINOLOGY MANAGEMENT

INFORMATION BOXES

Note: The authors of the Information Boxes are Sue Ellen Wright and Gerhard Budin, with the exception of Infobox 18, which is by Clove Lynch.

APPENDICES

The Handbook of Terminology Management

Introduction to Volume II

THE NATURE OF A HANDBOOK

In the Introduction to Volume I of this *Handbook of Terminology Management*, the compilers outlined the affinity of this work to handbooks treating, for instance, topics in engineering and the social sciences. Whereas Volume I covered basic, to a great extent timeless, topics in the field, Volume II addresses a range of issues involving pragmatic technical applications in terminology management.

The focus of the project on leading edge technologies has been the source of countless difficulties in completing the volume as planned. First of all, we initially (and perhaps naively) cast our net wide to encompass a broad scope of topics and applications involving the role of terminology in human language technologies and language engineering at a time when these fields were and are still undergoing unprecedented growth and innovation (see Infobox 31:887, *Language Engineering and Language Technology*). Not only did we find that the amount of space originally allotted to certain key topics was inadequate, we also discovered that it was exceptionally difficult to achieve closure in a number of areas, some of which were the direct locus of our own core research. Indeed, our involvement in the development of new technologies, for instance in terminology data interchange, has stood in the way of our timely concentration on this volume. This was notably the case with the chapters in Section 8.1, *COMPUTER APPLICATIONS FOR TERMINOLOGY*. The final approval and publication of ISO 12200:1999, *Computer Applications in Terminology — Machine Readable Interchange Format (MARTIF) — Negotiated Interchange*, has finally provided a sense of completion that underlies the spirit of Chapter 8.1.5, *Terminology Interchange*, but even as this volume goes to press, late-breaking developments in the evolution of terminology interchange for use in integrated computer environments have inspired last-minute changes to that chapter.

To further complicate matters, we found it almost impossible to synchronize the many different topics in this broad and constantly shifting field while at the same time trying to bring each article to a state-of-the-art status simultaneous with that of other articles. In the end, each article represents a certain cut-off point, a kind of snapshot of the topic in question, reflecting the conditions at a particular moment in time. We can only hope that the quality of the contributions will compensate for the impossibility of total temporal coordination.

As a result of the passage of time and evolving intentional aims, some of the annexes listed in Volume I have been revised or eliminated, and there have been a few

changes from the authors and titles announced in the earlier volume. We commend the reader to the *Table of Contents* in this volume for current information and apologize for any inconvenience or disappointments that these changes may have caused. At the end of our discussion in Volume I of the nature of a handbook, we noted that this text is intended to be a permanent companion of the working terminologist, a resource that is continuously *at hand*, designed to provide an ever-ready reservoir of useful basic information on the prevalent topics of concern to a broad range of users. We reiterate this goal with respect to Volume II, with the additional caveat that we hope to be able to maintain the currency of the more topical information included here by maintaining an updatable website at: http://appling.kent.edu/htmweb/htmweb.htm

TERMINOLOGY MANAGEMENT

In Volume I we defined *terminology management* as:

any deliberate manipulation of terminological information

We refer the reader to that volume for further treatment of the scope of this activity. Within the framework of language technology and language engineering, we want to emphasize that terminology management extends beyond *terminology database management* to include all aspects of terminology used as a component of information and quality management, as well as the role that terminology plays in document production and corpus management.

THE SCOPE OF TERMINOLOGY MANAGEMENT

The Organization of this Handbook

As noted in Volume I, the extensive topic range treated in this collection has dictated that this "book" be published in two volumes. The following overview provides a brief summary of the topics treated in Volume II.

VOLUME II: APPLICATIONS-ORIENTED TERMINOLOGY MANAGEMENT

Information Management

As information takes its position as a fundamental production variable (along with the classic variables, capital and labor), information organization and retrieval comprises a core concern. In chapter 6.1, *Computer-Assisted Thesaurus Management*, Roulin outlines the use of electronic tools for the creation of documentary languages used for indexing and querying documents. In 6.2, *Terminology and Indexing* Strehlow provides a brief overview of standard indexing practice and explores the critical principles of precision and recall in 6.3, *Terminology and Information Retrieval*.

Commercial and Industrial Applications

In one of the shifts in writing assignments that has occurred between the two volumes, Wettengel and de Weyer have combined forces in 7.1, *Terminology in Technical Writing*, to analyze ways in which technical writers can use terminological resources to model documents to fit the needs of their target audiences. In contrast to the traditional terminological focus on nominal forms, they stress the role that actions (verbs) play in good technical writing.

The growing role of terminology in industrial management is reflected by a series of articles on industrial operations and inventory control. In 7.2.1, *Terminology as an Organizational Principle in CIM Environments*, Wright describes the central function of terminology as the motivating force behind the assignment of data element names and definitions in computer-integrated manufacturing. Düsterbeck and Hesser (7.2.2) explore *Terminology-Based Knowledge Engineering in Enterprises*, recounting the cost savings that can be achieved through terminologically informed inventory control. Wright's *Terminology and Total Quality Management* (7.2.3) documents the role that parallel-text-oriented terminological research plays in ensuring accurate multilingual terminology. In 7.2.4, *Project-Integrated Terminology Management for Technical Writing and Translation*, Champe presents a case-study reflecting best practices for coordinating terminology research with commercial translation clients. In 7.2.5, *Software Terminology and Localization*, Corbolante and Irmler outline procedures for guaranteeing terminological appropriateness when adapting software to the individual market demands of different global locales.

Computer Applications for Terminology

The articles under the heading of *Database Management Methods* begin with advice on selecting *Terminology Database Management Programs* (TMSs) by Schmitz (8.1.1), who analyzes the essential features of terminology applications and advises the reader on selecting the right one for a given environment. In 8.1.2, *Data Categories for Terminology Management*, Wright discusses the distinguishing characteristics of the data categories used in terminological databases (termbases) and outlines the structure of the ISO 12620 standard on data categories for terminology management, explaining how the individual categories fit into the data model in different working environments. Wright goes on in 8.1.3, *Terminology Management Entry Structures*, to chronicle the evolution of terminological record and its modern manifestation as a terminological entry in a well-structured database environment, particularly in light of developments with respect to meta-models for representing cross-platform structural relationships. The article on *User-Specific Terminological Data Retrieval* by Nkwenti-Azeh (8.1.4) presents the prospect of multi-purpose, user-customizable terminology retrieval from large term banks. In the final article of this

section (8.1.5), Melby, Schmitz and Wright provide an introduction to *Terminology Interchange* with the intention of orienting the reader toward efforts to implement a universal terminology exchange format designed to facilitate the importation and exportation of terminological data using Standard Generalized Markup Language (SGML) and eXtensible Markup Language (XML).

Section 8.2, *Database Applications and Products* features three case studies that explore actual working environments within which terminology products are created and maintained. The three contributions in this section also document case studies, providing a snapshot view of a particular working environment viewed at a specific time in its development. Although it is impossible to maintain current updates on this kind of information, the compilers hope that the examples shown here will provide models for others in structuring their own tools and procedures. These case studies include *Preparing Multi-Volume Illustrated Terminological Dictionaries* (8.2.1, Le Néal) and *Practical Considerations for a Term Bank: Termium* (8.2.2, Hutcheson; referenced in the marginal notes as *Termium*). The case studies also include a chapter on *Globalization and Terminology Management* (8.2.3), which describes a highly complex terminological support system within a major corporate environment (Warburton). These topics lead naturally into the consideration of the role of *Terminology and Machine Translation* and the special requirements needed to fully integrate terminological information into machine translation systems (Vasconcellos, 8.3).

Corpus-related Applications addresses the question of automatic terminology generation in articles on the *Corpus Linguistics and Terminology Extraction* (Ahmad and Rogers, 8.4.1) and *Terminology Compilation: Consequences and Aspects of Automation* (Sager, 8.4.2). *Terminological Text Production* (Shreve, 8.4.3), and *Collocations in Sub-language Texts* (Heid, 8.4.4). These discussions underscore the relationship of terminological issues to issues involving discourse and texuality. The *Role of Specialist Terminology in Artificial Intelligence and Knowledge Acquisition* (Ahmad, 8.4.5) provides a structural overview of the importance of terminology in creating artificial intelligence and expert systems.

In 8.5, *Terminology Management Resources on the Internet*, Budin and Wright document the growing wealth of information accessible at sites on the World Wide Web (WWW) or available for downloading from the Internet using File Transfer Protocol (FTP).

Backmatter

The Information Boxes included in Volume II serve a similar function to those in Volume I and pertain specifically to the issues addressed in the second volume. The topics addressed here are generally ones that relate to several of the articles in the book and are more effective as shared resources than if they were treated separately

(or severally) in the other articles. Several of the annexes promised for this volume have actually been incorporated into the Infoboxes. In addition to the *List of Abbreviations Used in This Volume,* Volume II contains a listing of current standards and projects that are part of the work program of Technical Committee 37 of the International Organization for Standardization.

Volume I: Basic Aspects of Terminology Management

For the convenience of those who may not have Volume I of this handbook at hand, the introductory collection addresses the classic issues of terminology management, along with overriding practical and external factors, such as terminology training issues and copyright.

Fundamental Principles

The initial segment of Volume I addresses basic concerns involved in the documentation of terminology. *Term Selection* (Wright, 1.1) and *Term Formation* (Sager, 1.2) together provide insight into the most salient of all terminological activities: the identification and creation of terms used or to be used in a discipline and to be documented in terminology management projects. The chapter on *Graphic Representation* (Galinski and Picht, 1.3) supports the thesis that non-verbal means of representation occur in addition to the verbal forms that comprise the primary substance of most terminology collections, and that these forms of representation are proliferating in modern information environments.

Under the heading of *Concept Description*, de Bessé's article on *Definitions* (1.3.1) stresses the importance of complete terminological definitions, as opposed to some of the forms of so-called definitions used in other resources. In *Frames and Display of Definitions* (1.3.2), Strehlow uses frame theory to provide a new view of the components making up a formal definition. Dubuc and Lauriston (1.3.3) underscore the value of using *Contexts* to document the relationship of terms to concepts, the meaning of terms, and the way that they function in texts. Beginning with a basic explanation of the *Representation of Concept Systems* (Wright, 1.4.1), the discussion moves on to Eck, Meyer, and Skuce (1.4.2), who offer a *Systematic Concept Analysis within a Knowledge-Based Approach to Terminology*. Chapters treating the *Multifaceted* and *Multidimensional* aspects of concept classification (Kageura, 1.4.3, and Bowker, 1.4.4) explore efforts to overcome concentration on overly simplistic, two-dimensional systems for representing concept relationships.

Types of Terminology Management

The second major section of the book addresses the many different applications involved in terminology management.

The treatment of *Descriptive Terminology* management explores primarily non-normative terminological activities involving *Technical Translation* (Wright and Wright, 2.1.1), *Medical Science* (Lynch, 2.1.2), *Patent Language* (Lawson, 2.1.3), and the *Social Sciences* (Riggs, Mälkiä, and Budin, 2.1.4). These activities primarily involve documentation on the part of both experts and quasi-specialists (well-versed, non-expert terminologists) who prepare terminology products and online resources for use by both experts and lay practitioners, such as translators, technical writers, and students.

The articles pertaining to *Prescriptive Terminology* management, on the other hand, examine practices and methodologies that have evolved in the arena of normative terminology (Wright, 2.2. 1; Strehlow, 2.2.2; Schrade, 2.2.3; and Bowmann, Michaud, and Suonuuti, 2.2.4), primarily within the framework of national and international standardization, such as the International Organization for Standardization (ISO), and professional societies, such as the International Union for Pure and Applied Chemistry (IUPAC). Merritt's contribution (2.2.5) outlines resources for those working in the field of chemistry and provides a model for the development of highly structured nomenclatures.

Language Policy

Multilingualism in Terminology Management (Budin and Wright, 3.1) examines language policy as applied to countries and regions that must contend with multiple official and unofficial languages and even with the use of multiple scripts, which generates the need to cope with problems involving transliteration and transcription. It also addresses situations involving languages that are facing the challenge of developing technical terminologies in an orderly fashion in order to combat a confusing proliferation of synonyms and potentially ambiguous forms. *Language and Terminology Planning* (Humbley, 3.2) in the francophone environment documents both the history and the current development of French efforts in language planning.

Intellectual Property Rights

Copyright and Terminology (Galinski and Wright, 4.1) outlines the extent to which copyright issues affect terminology collections. In lieu of clearcut protection within the framework of national and international copyright conventions, the authors support the use of a *Guide to Terminology Agreements* as a useful tool for establishing safe and equitable contractual relationships among individuals and entities sharing and exchanging terminological data. Such agreements are especially valuable in flexible electronic environments where the merging and reuse of data gives rise to questions with respect to authorship and exploitation that as yet remain unclear within the framework of emerging trends in copyright protection.

Terminology Training

The chapter on *Terminology Training* (Picht and Acuña, 5.1) outlines the types of terminology training that are being practiced in the world today and proposes modular components that make up the subject matter of courses presented in different venues and for different intentional aims.

ORGANIZATIONAL PHILOSOPHY AND INFORMATION RETRIEVAL MECHANISMS

The topics treated in this book have been chosen with an eye to providing comprehensive coverage of each aspect of terminology management. They do not merely represent yet another collection of potentially unrelated or uncoordinated articles selected at random by their authors for presentation within a particular symposium or series collection. These subjects were carefully assigned to the individual authors, based on the assumption that they have distinguished themselves because of their work in specific subfields of the discipline. The articles have been fine-tuned, and a rich network of cross-references have been integrated into the text to facilitate hypertext-like movement within the book as a whole. These mechanisms are more clearly explained in the following section.

HOW TO USE THIS BOOK

Although the serious or zealous reader is cordially invited to read the book from cover to cover, it is not intended to present a one-time reading experience, but rather to be a permanent reference work that users will return to again and again, both to reconfirm prior knowledge and to familiarize themselves with unfamiliar areas of the field. To this end, we have provided the following devices designed for rapid navigation through the text in search of specific information:

- *Quick-search keywords*
  Marginal references that highlight topic concepts in the individual articles

- SEE

  Term Selection
  1.1.1:13

  SEE boxes cross-reference to other articles in this volume. Articles are indicated by chapter short name and number, followed by a colon and a page reference. For instance, in the above example, *Term Selection* is the title of a chapter, *1.1.1* the number of that chapter, and *13* the page on which the chapter

begins. In cases where the entire article is related to a given point in the text, the page reference is for the first page of the article. Where a specific passage in a chapter is referenced, the page number may reflect the page where the related topic is discussed.

- INFO

  Classification
  INFB 9:338

  INFO boxes act as pointers that guide the reader to *Information Boxes* treating basic concepts that are relevant to more than one of the chapters in the book; Infoboxes are indicated by a short name and number, followed by a colon and a page reference. In the example, *Classification* is the short name for Infobox 9: *Classification and Documentary Languages*, which begins on page *338*.

- A detailed, structured subject index designed to allow easy access to topics discussed in the articles, together with an authors' index for easy reference to authors cited in the book

In addition to our many authors, we wish to acknowledge the contributions of our assistants Isabelle Cannas, Anke Förster, Yuthika Prabhu, Elisabeth Skorianz, and Anne Beinchet, as well as the work of our indexer, Michelle Foss. Thanks also go to our careful first readers, Bertie Kaal, Leland D. Wright, Jr., and Juan C. Sager, as well as Kara Warburton. Needless to say, we also want to thank Aban Budin and Leland D. Wright for feeding us well and putting up with the chaos generated by such a massive project.

Sue Ellen Wright
Kent, Ohio, U.S.A.
July, 2000

Gerhard Budin
Vienna, Austria

Section 6

Information Management

Section 6.1

Computer-Assisted Thesaurus Management

CORENTIN ROULIN

1 INTRODUCTION

1.1 Aim and Contents

This chapter provides:

- an overview of how computers can assist thesaurus management (creation, production, and maintenance) and
- assistance in evaluating, choosing, customizing, and implementing a software tool for this purpose

thesaurus use

thesaurus management

After distinguishing between thesaurus management and thesaurus use, the chapter examines data and data structures related to handling thesauri in the form of computer files. Thesaurus management systems must obviously provide functions that allow users to consult the thesaurus, but they must also provide management features for building the thesaurus, updating it, and producing printed and electronic editions. Finally, system designers must provide for an effective interface to the "ideal" thesaurus management system described throughout the course of the chapter.

1.2 Thesauri

documentary language

A thesaurus is a *documentary language,* i.e., the "common language" used in a documentary system. Thesaurus managers use it for representing (*indexing*) the content of the recorded documents (or any type of item that is to be described using thesaurus descriptors in an information system), and users employ it to query the subject they are interested in and thus for *retrieving* relevant recorded documents. In this text, the term *document* is used to designate such an item.

syntax

As a documentary language, the *vocabulary* of a thesaurus consists of terms selected from natural language, as opposed to the artificial codes used in a classification scheme. The *syntax* of such a language is quite basic in most documentary systems. In indexing,

the juxtaposition of indexing terms (known as *post-coordination*) is generally the only way to represent both the specific aspects of the subject treated in a given document and the extent of its content. In order to retrieve documents, users enter indexing terms or groups of terms together with Boolean operators (AND, OR, NOT). Functions can also be provided to take advantage of the *a priori* relations that give structure to the vocabulary of a thesaurus.

conceptual structures

A thesaurus is more than a mere list of terms. It is a semantically structured set of concepts. The vocabulary is organized according to relations so that it will be easier to understand, survey, and use in a way that emphasizes the meaning of the terms that comprise it. To ensure its efficiency, a given thesaurus covers a limited number of fields of knowledge in detail and others more superficially. This specialization is marked in the choice of the specific terms introduced (by "pre-coordination", i.e., the introduction of compound terms) and by establishing the relations between these terms.

multilingual thesauri

A *multilingual thesaurus* allows access to common documentary data sources for users who speak different languages. Terms are selected from more than one natural language for each concept present and the terms are treated as linguistically equivalent. Consequently, the semantic relations are, as a rule, strictly identical from one language version to another since they actually link the concepts behind the words.

1.3 Thesaurus Management

thesaurus use

If using a tool includes being able to maintain, improve and customize it according to specific needs, thesaurus management can be viewed as one component of thesaurus use. However, using a thesaurus primarily consists of indexing and retrieving documents, i.e., assigning descriptors to documents and expressing queries using the same vocabulary. User-friendly performance of these two tasks in a given computerized information system is essential, but this function is outside the scope of this chapter.

user and management needs

Both thesaurus users and managers need to consult the thesaurus, which implies searching for terms, displaying concepts, and navigating among these concepts based on the relations programmed into the system. Manager, however, have different needs from users, notably because managers need to view the thesaurus as a dynamic system—one could say as a living object—whereas

users want to rely on a stabilized version of the documentary language. This is especially true when a thesaurus is used by more than one group of users, i.e., in more than one information system, which introduces numerous specific requirements for thesaurus management.

1.4 Computer Assistance

paper vs. computer-based systems

Despite today's prevalence of computer-based systems, it is still possible to envisage thesaurus management using pencil and paper only. Experience shows that this means using several pencils, an eraser, a large amount of paper—or, preferably, index cards—and a card file. Thesauri are paper-hungry. This chapter, however, concentrates on software tools designed for professional thesaurus management, both those that are available on the market and those that could be built in the future. It excludes discussing non-specialized tools, such as word processors or spreadsheets, which are sometimes used for this purpose.

1.4.1 *Specialized vs. integrated tools*

integrated tools

Thesaurus management has its own specific requirements that are irrelevant for users who have to index and retrieve documents or that are not even compatible with the requirements of these users. Information systems, including thesaurus management modules that allow direct (interactive) modifications to the thesaurus, are designated as *integrated tools*. When using these tools, it is more difficult, or even impossible, for thesaurus managers to perform certain operations such as:

- introducing modifications on an experimental basis
- reviewing previously introduced modifications and systematizing the collection of update data designed to inform users
- revising indexations resulting from previous modifications as a separate task
- producing printed editions on small scale-systems during normal use time or launching any other memory-intensive computer operations that are likely to overload the system—thesauri are especially memory-hungry

However, integrated tools have advantages. They provide:

- immediate availability of corrections and updates

- immediate availability of usage statistics
- identical functions and interface for both managers and users

specialized tools

Nevertheless, a *specialized tool*, i.e., a separate software module for thesaurus management, is recommended where there is more than one user group or target information system, for such cases usually require the isolation of update work in order to:

- control the management of successive versions and reserve equal treatment for all user groups
- formalize updating data to be transferred to target systems, i.e., separate user systems that may demand different input formats

avoiding drawbacks

The possible drawbacks of using specialized tools run parallel to the advantages listed above for integrated tools. They can be avoided by:

- ensuring a straightforward update process from the management system to the target user system(s)
- ensuring straightforward dataflow for use-related statistics from user systems to the management system in cases involving completely separate systems
- using the same or compatible software for management and for user functions

1.4.2 *Data Modeling*

common data structures

Defining a computerized system includes describing the data to be handled and their structure, as well as defining the different functions (creation, update, report generation) the system must perform on these data in order to achieve its intended management task. As far as data and data structures are concerned, all thesauri share common characteristics. For instance, all thesauri include preferred and non-preferred terms to express concepts and define semantic relations between concepts, but many thesauri also have features of their own. For example, a given thesaurus may distinguish between two types of hierarchical relations—namely both specific-generic relations and partitive relations—where others do not include such a refinement. This chapter takes as a starting point the elements and relations described for multilingual thesauri in the relevant international standards (ISO 5964), with the exception of "facets".

1.4.3 *Functions*

overlapping functions

As explained above, users and managers require two sets of overlapping functions. Figure 1 illustrates this overlap and provides references to the corresponding parts of this chapter that deal with each group of functions. Indexing and retrieval are discussed only briefly.

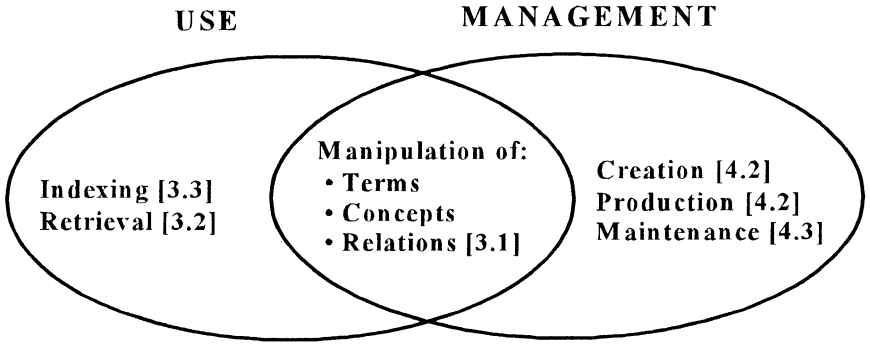


Figure 1: *Thesaurus Use and Management Functions*

customized systems

In some cases, especially when creating custom-tailored information systems, additional modalities are possible. Not only are thesaurus manipulation (consultation) and management treated as separate functions. The operation that links selected concepts to documents (indexing), query formulation, and retrieval tools are also separate functions. Quality is the most important issue in this regard, notably the reliability and the speed of the interface between the thesaurus consultation module and each of the other modules (thesaurus management, document indexing and document retrieval).

1.4.4 *User-Friendliness*

system specifications

Although user-friendliness is often a hidden dimension in the lists of software evaluation criteria, it is a major factor in deciding to select any given system. This chapter does not treat user-friendliness in detail, but appropriate references will be made whenever this dimension is relevant. In some cases, the corresponding requirements are mandatory ones. To cover all the other cases, it would be beneficial to isolate this criterion in the list of selection

specifications in order to give it appropriate priority, especially when customizing a system, where clearly stated guidelines will help when negotiating with the developers.

2 DATA AND DATA STRUCTURE

2.1 Conceptual Data Modelling

2.1.1 *Introduction*

data modelling levels

Software development techniques distinguish among these successive levels used to describe the data to be manipulated by an automated system according to the degree of conceptualization inherent in these descriptions. The *conceptual level* most nearly approximates the way users look at the problem: data are described in terms of concrete or abstract objects that are manipulated in the domain under consideration according to the relations that exist among these objects. The *logical level* uses a particular class of software tools to integrate constraints proper to computerized data management, e.g., the description of data to be handled with a relational database management system will have to be made compliant with the various "normal forms". The *physical level* most closely reflects the actual requirements of a given system in a given hardware and software environment. Performance is then the major concern of the system developer, who has to ensure that the planned system will be able to manipulate the data efficiently in order to achieve user objectives. Transforming a conceptual data model into a description of files to be implemented at the physical level is sometimes designated as a *degradation* process. For instance, a typical element of this degradation involves allowing for some data redundancy.

degradation

conceptual level

In this section, data to be manipulated for thesaurus management are described at the highest conceptual level in order to provide the reader with a description that remains valid whatever tool is used and that allows comparisons between different degradation approaches (see 2.2.1 below). The following section describes some typical implementation options and draws attention to the corresponding potential drawbacks for the thesaurus manager. The last section of this chapter is devoted to practical considerations with respect to the data as they have to be actually manipulated by the computer.

2.1.2 *Data Elements and Relations*

Thesauri can be described as systems made up of sets of data elements and of the relations established among those elements. These relations constitute a structural description of a conceptual view of the system in which significant subsystems are progressively delineated.

descriptors

2.1.2.1 *The Concept Subsystem.* The concept subsystem is made of the three basic elements—descriptors, non-descriptors, and scope notes—and of the relations that link them. The *descriptors* are the authorized indexing and retrieval terms. They consist of single or compound terms, generally in the singular form. Adjectives and verbs are excluded. *Non-descriptors* are entry terms that cannot be used as indexing terms. They are synonyms or quasi-synonyms of the terms chosen as descriptors and are provided to help the user find the most appropriate authorized indexing by referencing descriptors.

non-descriptors

By linking each non-descriptor to the descriptor to be used instead, the semantic equivalence relations form bridges between the vocabulary of natural language and that of the artificial language. Multi-equivalence relations may also be used to link one non-descriptor to two descriptors either to be used together (post-coordination guidance) or between which the user is invited to choose (in the case of polysemes).

multi-equivalence relations

Scope notes are short texts that either give a definition for, or state restrictions and give advice on the use of, given descriptors. Application relations link each scope note to the descriptor to which it applies. Some thesauri distinguish between scope notes, definitions, and "history notes", which contain information on how the concept was expressed in previous versions of the thesaurus.

scope notes

In a monolingual thesaurus (or in a given language version of a multilingual thesaurus), the intended *concept* can only be grasped correctly by considering together the chosen descriptor, its non-descriptors, and any scope note that might have been attached to it.

Figure 2 shows that some elements belong to "classes" (namely *terms* and *texts*), that is to say, they share common formal characteristics. These two classes are especially important in a multilingual thesaurus: language identifier relations link each term and each text to a single language, which allows for the distinction

multilingual thesauri

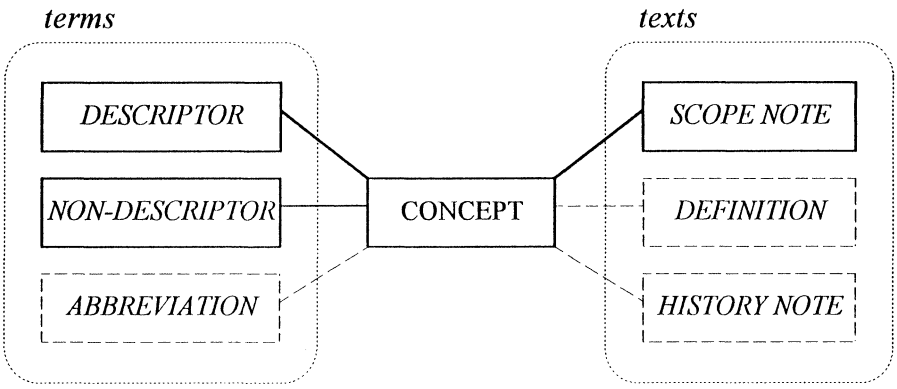


Figure 2: *Element Classes*

among the different *language versions* as subsystems. Typical attributes of terms and texts are their type, language, and wording. For management purposes, it is also necessary to record and maintain information on proposals and decisions taken related to each element.

2.1.2.2 *Structuring relations and subsystems. Hierarchical relations* are established between pairs of concepts where one term (referred to as the *narrower term*) can be considered as subordinate to the other (*broader term*). The distinction between specific-to-generic and partitive relations is usually not made in thesauri. Some thesauri allow polyhierarchical relations in which one concept can be associated with more than one generic concept.

hierarchical relations

The purpose of this essential means of structuring a thesaurus is to help users explore the vocabulary in a systematic way. It can also be exploited in retrieval mechanisms. *Hierarchical chains* are subsystems made of concepts linked together by hierarchical relations. *Associative relations* are established between concepts designated as possible alternative indexing terms, which form complementary paths for exploring the vocabulary.

hierarchical chains

associative relations

Inclusive relations link each hierarchical chain—and, as a consequence, each concept—to a super-set that is generally defined according to a classification-like approach. These relations are sometimes organized into several levels. The *domains* subsystems (sometimes designated as "fields" or "microthesauri" and

inclusive relations

identified by codes or titles) thus form the highest levels of the semantic structure of the vocabulary.

Figure 3 emphasizes the relations (hierarchy, association and inclusion) because, for management purposes, it must be possible to manipulate them as such and to record and maintain information about them, such as proposals and decisions. The "hierarchy/order" relation is used to manage domains and sub-domains and to order them according to a logic different from mere alphabetical order.

2.2 Implementation Options

2.2.1 Degrading the Model

One customary approach to implementation is to consider the concept as a single object and to manage it in a single file made of records with fields to encode descriptors, non-descriptors and scope-notes. This option leads to the following problems:

single-object concepts

- It is necessary to limit the number of non-descriptors and to determine their maximum number beforehand.
- It is impossible to attach a given term to another concept without re-typing it.
- It is necessary to scan all "concept" records to perform functions involving all terms or certain types of terms, such as tracking duplicates among descriptors and non-descriptors or extracting all modified descriptors.

managing relations

Managing relations is an important aspect, especially in non-relational database management systems. Ways of implementing "one to many" and "many to many" relations will determine how easy it will be to perform operations such as linking a given concept and all its narrower terms to a new generic concept, removing all relations for a deleted concept, and restoring a previous situation.

degradation options

For the sake of performance, some degradation options may be chosen for thesaurus use. For example, in some cases it can be difficult to follow the line from a given concept to the domain(s) to which it belongs via the hierarchical relations to the most generic concept(s) and then from this (these) top-term(s) to the domain(s). If this is the case, the code or title of the domain(s) can

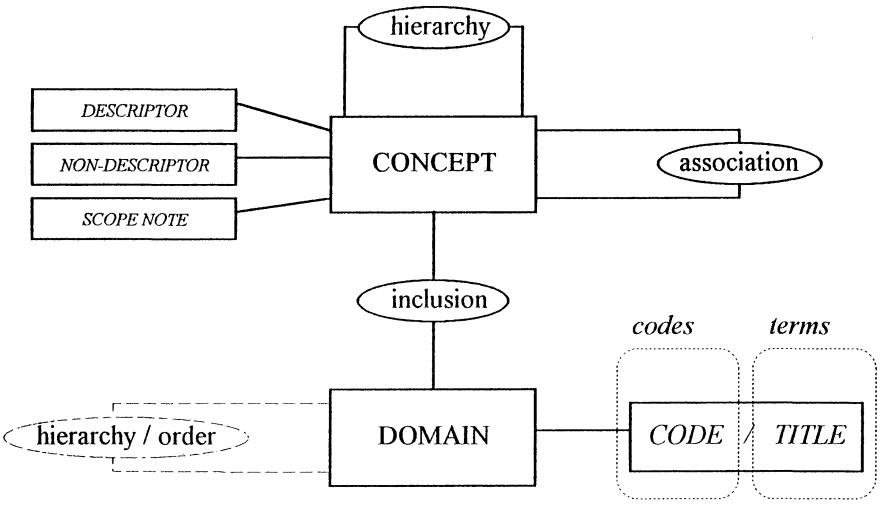


Figure 3: *Thesaurus Relations*

conceptual models

be encoded in the concept record so that it is unnecessary to have to perform this search every time the information is needed. This approach can, however, lead to significant redundancy that will become problematic if a domain title or code is modified. Nevertheless, thesaurus management will be more efficient if all the elements and relations handled are implemented as separate objects. In other words, the closer the implementation is to the conceptual model, the safer. The first step of computer-assisted thesaurus management involves establishing the conceptual data model that corresponds to the thesaurus being designed.

2.2.2 *Off-the-shelf vs. Custom-tailored Systems*

software selection

Numerous software programs designed for managing fully standard-compliant thesauri have been coming on the market. Selecting an existing product requires a carefully conducted assessment process. The more numerous the specific characteristics required for the application, the more attractive a custom-tailored approach begins to look. Needless to say, a customized product will inevitably be more expensive. The corresponding cost-benefit analysis should include the relative *costs* of:

cost-benefit analysis

- using an available software development team that would be capable of creating a tool from scratch
- taking advantage of parameter configuration options offered in an existing software program (Note that the question of who does what at what price always has to be answered very precisely.)
- benefiting from custom adaptations that could be proposed by the provider of an existing tool

qualitative considerations

This analysis should also include purely *qualitative* considerations such as the ability of the envisaged software development team to deal correctly with textual data and, more generally, with the complexity of the thesaurus management environment, which should not be underestimated. Another option is to outsource the whole thesaurus management function (see Section 5).

2.3 DATA MANAGEMENT

2.3.1 *Data Elements*

limitations

2.3.1.1 *Length of Textual Data.* Some software systems limit the length of terms, which is quite efficient if the users are required to type in the descriptors when indexing documents or formulating queries, but which is irrelevant if this is not the case. Ideally, it should be possible for terms to be as long as necessary to take full advantage of one of the basic characteristics of thesauri as documentary languages, which is the ability to use words instead of codes to express concepts. Depending on the thematic field covered by the thesaurus, the typical limitation to between 30 and 40 characters may be unacceptable. This holds true for texts as well, e.g., in the scope notes. When textual data longer than terms are stored in special text-fields, the retrieval aspect also has to be especially examined, e.g., is it possible to perform quick searches on words in these data?

multilingual character management

2.3.1.2 *Character Sets.* The ability of a given software program to satisfactorily handle textual data in a given language or group of languages is essential, especially in the case of multilingual thesauri, but also in the case of a thesaurus in one "rare" language, i.e., in any language for which no fully integrated solution for textual data management is as yet widely available. Beyond the thesaurus management system itself, the software and hardware

environment in which it is to be used (database management system or programming language, operating system, keyboard, display, and printing devices) will determine if and how efficiently textual data can be typed in, recorded (using one or more than one byte per character), displayed on the screen, searched, sorted alphabetically, or printed. (Each element of this list is equally important). Solutions for one language at a time exist for most languages. Problems appear, however, when it is desirable, for instance, to display several character sets on screen at the same time. In this case, compromises are likely to be necessary. Before accepting any given solution, one should consider:

INFO

Unicode
INFB 20:861

- which form of the thesaurus is likely to be the most used (the printed or the on-screen version)
- whether the use of upper and lower cases—together with accented letters—is important, which may differ considerably according to the language
- whether different formats should be necessary, e. g., a complete character set for the printed version, but upper case only for the target information system

2.3.1.3 *Numerical and Coded Data.* A thesaurus management system has to allow for the manipulation of numerical data as well. The thesaurus manager needs statistical data that give information on the characteristics of the thesaurus as such and on the frequency with which specific concepts are used in indexing and retrieval. As far as coded data are concerned, a thesaurus management system should offer all functions generally available for such data, such as choosing from a pull-down menu that presents the limited set of permissible values (e.g., language and type of a term, or the status of a modification).

statistial data

2.3.2 *Relations*

Whatever the method of managing relations at the physical level, a thesaurus management system must provide the manager with all the necessary functions required to take and keep track of decisions on relations, to easily manipulate reciprocal and transitive relations, and to perform consistency checks. *Modifications of relations* include the addition, deletion, and modification of linked objects. Quite often, a single modification has consequences that affect more than two objects, which means that the thesaurus

consistency checks

modifying relations

manager needs a system that requires minimum typing and provides maximum capability for automated processing and data control.

reciprocal relations

Thesaurus managers also need to be able to change their minds, which supports a preference for approaches that isolate relations as objects, allow contradictory relations to co-exist, and so enable precise control of relation status. As far as *reciprocal relations* are concerned, it is mandatory that the system take them into account automatically, i.e., if A is said to be a narrower term of B, the relation indicating that B is a broader term of A must be automatically set.

transitive relations

Transitive relations include, for instance, the "inclusion" relation whereby if a "top-term" is attached to a domain, then all its (direct and indirect) narrower terms also belong to that domain (sometimes called inheritance). The main concern of the thesaurus manager is that the consequences of a given modification must be taken into account so as to make all the consecutive modifications anywhere else in the thesaurus available immediately, and, if the case arises, to generate a report of inconsistency problems before any further modification is introduced.

rules

Rules are applicable to keep the structure of a thesaurus, which is sometimes referred to as a "semantic network", consistent. For example, it is admitted that a given term cannot be at the same time, directly or indirectly, both a narrower and a broader term of the same term, which would lead to unmanageable endless loops in the network.

polyhierarchy

A distinction has to be made here between standard rules and rules specific to a given thesaurus, and also between "mandatory" rules, the violation of which will not be accepted by the system and which will immediately generate an error message, and rules whose violation will only make the system generate a warning to the manager, possibly not online but in the form of a printed report. For example, polyhierarchy might be admitted in some thesauri and not in others. If it is admitted, it might be recommended to limit its use to concepts that are at the lowest level in the hierarchical chains, i.e., concepts to which no further narrower concepts are attached.

Thus, thesaurus management functions—and this is especially true for ready-made systems that incorporate *a priori* consistency rules—should include options for:

- tailoring the mandatory level associated with each rule that is implemented
- implementing new rules

3 User Functions

3.1 Searching for and Manipulating Concepts

3.1.1 *Retrieving Terms*

3.1.1.1 *Formulating Queries.* All data types should be searchable, including management data types such as dates. Since the basic elements of a thesaurus are terms or are made up of terms, the search functions of a thesaurus management system should include options to formulate queries with search criteria based on words and phrases, as well as the ability to:

thesaurus search functions

- search for exact wording (which is definitely enhanced by having access to the computerized inverted file or index in the form of a "word wheel") or for the occurrence of word(s). In case of words, the system should allow for the indication of definite order (or not) and of proximity operators.
- right, left, and medial truncation(s) (any number of characters) and wild cards (definite number of characters) on any word

Moreover, it should be possible to formulate queries that are limited to either:

- a given type of element ("search descriptors that ... [search criterion]" or "search non-descriptors ..." or "search scope notes ...") or not ("search elements of any type that ... [search criterion]")
- a language version (or not), taking into account that accents are essential in some languages and not in others

3.1.1.2 *Query Results.* The first result expected by users is to know whether the search was successful or not. If it has been, they want to know the number of hits, and if the query has been made on more than one type of element, they need to know the distribution of hits among the different types so that they can choose which element(s) they want to manipulate next. What the end user (i.e., the one who indexes and retrieves documents) needs is the concept. From the word specified, it is not necessarily useful to

hits

user needs

stop on the element found (term or text) before accessing the desired concept.

manager needs

The situation is different for thesaurus managers, for they look at all elements as distinct objects. That is why, if the terms and texts are implemented as distinct objects for both the end user and the manager, it must be possible to disconnect the mechanism that automatically proceeds to the next step (e.g., from the retrieved scope note or non-descriptor to the corresponding descriptor or concept, or from the domain title to the top-terms).

user expectations

3.1.2 *Displaying Concepts*

data display and access

Here again, needs may differ for the end user and for the thesaurus manager, but in case of a multilingual thesaurus, both will first want to choose a working language that will determine which language version is used with respect to display options (without preventing access to all data). End users primarily need to identify the concept presently available for further manipulation (the current concept). In most situations, providing them with the corresponding descriptor in the working language will be sufficient. In other cases, they will want to consult all the information available about the current concept, which may be partly available directly in the current record and partly distributed over linked objects. This information should be displayed in a manner that respects the following principles:

data principles

- All data in a record should be shown, or accessible at a single keystroke, and coded data should be made explicit (e.g., "English" rather than "language 2").
- All existing relations should be shown and made explicit (e.g., "A and B are the narrower terms" rather than "there are two narrower terms"), taking into account the working language.
- Format should be based on that of the "reference" presentation of the printed thesaurus, even if screen and paper are not comparable. Computer interactivity can be used to compensate for the limited screen size. The printed version is a good reference because it is familiar to users and has probably been designed to be as user-friendly as possible (see 4.2.)

graphical presentation

As far as graphical presentation is concerned, a graphical display of hierarchical chains parallel to the one provided on paper will be much appreciated, provided that computer screen generation time

is not excessive, and especially if all actions normally possible remain available in this context. The thesaurus manager's specific needs may be derived *mutatis mutandis* from the preceding paragraphs taking into consideration that they are interested in:

- not only concepts but all elements and relations as distinct objects, each one having its own display format
- management data not necessarily available to end users
- possibilities for tailoring display formats, including options for explicitating relations, such as showing status, e.g., "A and B are existing narrower terms, C is a new one, D is a proposed one".

3.1.3 *Exploiting Relations*

network navigation

Relations are essential to thesauri, and computers offer the possibility of navigating through the relational network and exploitation techniques that go far beyond what is possible on paper. Desirable functions can be described in a few words, but user-friendliness, which may be the most important aspect, is certainly the most difficult to describe. It must be possible to go from any element to any other element linked to it by any type of relation with a single keystroke or mouse click. (This operation is so frequent that having to press two keys is one too many.) It should also be possible to keep track of the path that has been followed and so to return to any previous position in the network.

3.1.4 *Manipulating Sets*

accumulating sets of objects

What end users and managers need when consulting their thesaurus is not always a single object (such as one concept) but quite often a set of objects (e.g., a set of concepts) that is of interest for their current task. Such objects are encountered in the course of performing searches on terms and navigating along relational links. One way to manipulate sets involves the possibility of picking up these objects as one works along without stopping the thesaurus consultation and losing the context.

identifying sets

A second way to manipulate sets involves selecting sets of elements linked to the current one, for example, all narrower concepts of the current concept (on one, several, or all levels of specificity) or all top-terms included in the current domain. An interesting, more sophisticated form of this mechanism can be applied not only from a single element but from a set of selected

elements (the "current set"), for example by selecting all concepts related to the ones in the current set. In using such a mechanism, users should be queried so that they can specify which logical operation shall be applied. In the example, they might be asked if the selection of concepts is related to *any* concept or to *all* concepts in the current set.

Boolean operators

When selected sets are available, the third way to manipulate them is to perform Boolean operations using techniques like those offered in documentary systems where a history of the previous searches is available, for example, selecting all concepts having "vitamin*" in the wording of the descriptor AND those having this word in the scope note (but) NOT those belonging to the domain "chemical products".

3.2 *From Concepts to Documents: Retrieval*

document records

One type of element among those that can be linked to concepts could be document records, in which case the generic mechanisms described in the previous section can be used directly for document retrieval purposes. If thesaurus consultation remains a separate module, it may be useful to automate some of the concept manipulations that are needed for formulating and tailoring "search strategies" and that imply the semantic relations linking concepts in the thesaurus.

posting

Such manipulations include "posting", which in this context means collecting concepts that are linked to a given concept by a specified type of semantic relation:

- "Generic posting" involves selecting the broader concept(s) of the current one.
- "Specific posting" involves selecting the narrower concepts, in which case it is necessary to be able to specify the number of levels of specificity to be taken into account.
- "Associative posting" involves selecting related concepts. Another form of this kind of posting involves selecting concepts that are "related via hierarchy" (or "siblings"), that is to say, concepts that have the same broader concept as the current concept.

3.3 *From Documents to Concepts: Indexing*

Indexing can be seen as "translating" the natural language of documents into a documentary language. Some automated systems

offer textual data analysis tools that tend to transform (reduce) the natural language of titles, abstracts, or full texts into sets of single and compound terms that can be compared to those appearing in the thesaurus used. Such a transformation implies *normalizing* the terms encountered in natural language by applying the same rules as those used during thesaurus elaboration, such as restricting terms to nouns, to the singular form, to a limited character set, etc. This process may also include the use of weighting mechanisms.

documentary language

Thesauri can also be enriched to favor comparisons with natural language by introducing more numerous non-descriptors as "lead-in" terms to fill the gap between intellectual and automated analysis of text content. Such non-descriptors range from synonyms that may be useful for guiding indexers to mere plural forms or spelling variants that are useful for the computer only. A more basic, though practical, function allows indexers to interactively select terms in available textual data and to request a check of their selection against thesaurus contents. In any case, the result of automated processes must be proposals for concepts and indexing statements that will be validated by the human indexer.

automated processes

human text analysis

4 Management Functions

4.1 Thesaurus Creation

4.1.1 *Data Collection*

Computers can certainly assist thesaurus management from the very early stages. If the corpus of documents that the corpus is designed to index (or a representative sample of this corpus) is available in electronic format, systems especially designed for this purpose can be used to extract terms from this source of "raw" data. In this case, attention should be paid to:

SEE

Term Extraction 8.4.1:725

- the nature of quantitative information provided (Basic information units include the number of occurrences for a term, and further ranking can take into account the position of occurrences in the analyzed texts, such as "in title", "in abstract", "in conclusion".)
- the nature of extracted terms, which should not just be single terms, but compound terms as well, and could be made available in a "normalized" form (see 3.3)

- the availability of information on co-occurrences of extracted terms
- the possibilities for customizing or fine-tuning the tool (regarding the extent of the corpus, extraction strategies, ranking techniques, etc.) so as not to be overwhelmed by a large amount of unmanageable data

structured resources

There might also be an opportunity to take advantage of existing sources of processed data, i.e., structured or controlled vocabularies such as dictionaries, glossaries, or any type of lists of terms, the contents of which include added semantic value relevant to the scope covered by the planned thesaurus and to the context envisaged for its use. Existing documentary languages—including classification schemes—are a particular example for such resources and may provide a starting point for building a new thesaurus based on portions of an existing thesaurus (or thesauri).

data importation and migration

In all these cases, the main issue is the interface or the import tool: how will available or extracted data be loaded into the thesaurus management system where it has to be used? This question is also relevant to the case of migrating from one thesaurus management system to another: how and at what cost will the data be loaded into the new system? This issue involves two major considerations:

exchange formats

- There is a need to agree on data exchange format(s) (see 4.2.2.).
- The thesaurus management system must be able to deal with data that might still be (partly) unstructured (see 4.1.2).

projecting costs

Even when some automatic data collection is possible (as described above), and certainly when this is impossible, thesaurus creators have to face the problem of manually encoding a large amount of data. Concern for the ease of (massive) input may look like addressing a "one-off" problem but, given staff-related costs, it might be a critical question to answer how the large number of words that will make up the thesaurus will be encoded. One option may be to use a customized, simplified encoding module with the appropriate structuring functions in order to achieve the right balance between input speed and the completeness of data required for management purposes.

4.1.2 *Structuring*

manipulating relations

It must be possible and even easy to manipulate relations (see 2.3.2). The early stages of thesaurus creation involve more specific requirements with regard to manipulating sets of elements and performing related actions on them. Thesaurus creation may require intermediate development stages where the nature of some terms (terms which are neither descriptors nor non-descriptors yet) or the nature of some relations (e.g., "grouping" of terms into a domain without proper inclusion relations) are still provisional. Structuring functions could include mechanisms such as:

structuring functions

- linking a grouping of terms to a "pseudo-concept" (a concept to which no attributes have yet been assigned)
- choosing a preferred term from among grouped terms and automatically assigning the appropriate thesaurus status (the chosen term is the descriptor, all the others are nondescriptors)
- grouping concepts to be precisely structured later

ongoing management

Structuring functions that are also useful during ongoing thesaurus management include:

- in case a concept with broader and narrower concepts is deleted, automatically modifying relations by hierarchically linking the "orphans" to the concept that serves as their "grandfather"
- selecting or listing all polyhierarchies, multi-equivalences, and duplicates

4.1.3 *Enrichment*

checking mechanisms

Once the structure of the thesaurus has been stabilized, its manager concentrates more on completing and enriching thesaurus contents. At this stage, manual input is inevitable, but intellectual work becomes much more important and time-consuming than mere encoding. In the case of a multilingual thesaurus in particular, managers will want to automatically check the completeness of contents, for instance by identifying concepts with missing descriptors (in a given language or in just any language).

processed data

For establishing language equivalents or generating scope notes (or definitions), managers might also benefit from the availability of multilingual sources of "processed" data for automatically loading such terms and texts (see 4.1.1). This phase of thesaurus

creation is also when the manager starts using available statistics on thesaurus characteristics (see 4.3.5).

4.2 Thesaurus Production

4.2.1 *Printed Formats*

4.2.1.1 *Presentation Types.* ISO 5964 provides very extensive specifications for the possible design(s) of printed editions of thesauri. They distinguish between alphabetical, systematic, and graphical presentations, and state that if there is more than one type of presentation, one of them must contain all available information about each concept (the "reference" presentation). A *structured alphabetical presentation* is usually established as this reference presentation. It lists all descriptors and nondescriptors in alphabetical order, with each term accompanied by the following data:

structured alphabetical presentation

- nondescriptors: the descriptor to be used instead
- descriptors: domain, language equivalents, scope note, nondescriptors, broader terms (on all levels), narrower terms (on all levels), related terms

This type of presentation is not practical when users want to look at the contents of the thesaurus according to its semantic structure. This is why the structured alphabetical presentation is usually supplemented by a *systematic presentation* or by a *graphical presentation*, both of which are more "use-oriented" in that they are organized according to domains and display complete hierarchical chains (sometimes with associative relations) and usually contain descriptors only. If it is impossible to incorporate language equivalents in the "reference" presentation, they can be given in a separate *multilingual alphabetical presentation*. Some thesauri also provide users with a means to search for terms on the basis of words as they appear in the form of a *permuted alphabetical presentation*.

systematic presentation

graphical presentation

A thesaurus management system must allow for producing the printed editions that are most desirable for users of the thesaurus, i.e., the system must not impose or limit the print formats. Different systems offer different solutions for printing:

print formats

- Some systems have a given number of available print formats that cannot be changed.
- Some systems offer options for reconfiguring print format parameters.
- Some systems include or give access to a report generator that allows for full reconfiguration of existing format parameters and for creation of new ones.

print output

Viewed from a technical perspective, actual printing options range from direct control of the printer by the thesaurus management system to production of tagged text files to be processed by word-processing or publishing software. These variations depend on the programming language or database management system used, which may include more or less sophisticated publishing features. When selecting a system, planners need to consider the availability and flexibility of these options, along with some of the specific questions discussed below.

alphabetical presentation

4.2.1.2 *Alphabetical and Systematic Presentations.* To produce alphabetical presentations, it is necessary to be able to *sort* a large number of terms and, in some multilingual thesauri, to do so according to sorting orders that differ from one language to another. Given that it might be impossible to rely on default sorting options implemented in low-level software layers, it is recommended that a thesaurus management system have its own sorting tools or that it allow for customizing the sorting tools used, or both. The question of the character set(s) handled has to be examined in connection with this point (see 2.3.1).

permuted presentation

KWIC Index

Producing a permuted alphabetical presentation introduces specific problems: the terms listed have to be sorted not only according to their complete wording (i.e., their spoken order), but also based on each "non-empty" word that appears in the term. This implies maintaining a (language-specific) list of "empty words" (or "stop words"). In some agglutinate languages, this also means configuring the permuted presentation so that it is possible to split some words into their meaningful components and to produce the appropriate page layout to display these lexical components (keyword in context or KWIC; see Figure 4).

	Audiovisuelles Programm
Ausbildung der	**Ausbilder**
	Ausbildung der Ausbilder
	Ausbildungsprogramm
	Ausbildungsrichtung
	USE Studienrichtung
	...
Audiovisuelles	**Programm**
Ausbildungs~	**programm**
	...
Ausbildungs~	***richtung***
	USE Studienrichtung
Studien~	**richtung**
	...
	Studienrichtung

Figure 4: *Keyword in Context Representation*

For the production of structured and systematic presentations, some systems limit the number of levels of specificity that can be listed recursively (together with alphabetical order between terms of the same level). Depending on the thesaurus, such a limitation may be unacceptable. It is especially useful in multilingual thesauri in particular to be able to customize parameters for relation indicators.

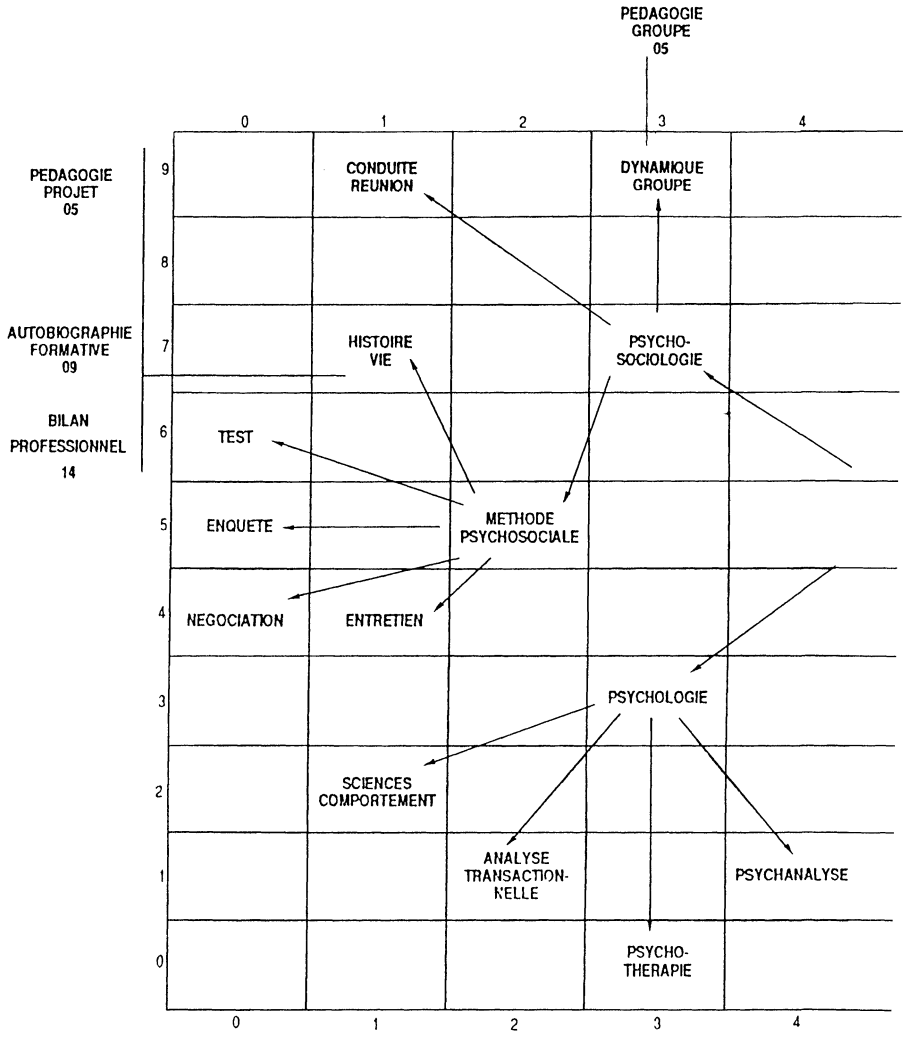
4.2.1.3 *Graphical Presentations.* Off-the-shelf software solutions that provide integrated graphics capability are rare. What would be useful for thesaurus management would be a system where modifications introduced in a graphic would be taken into account by the system and "translated" into modifications in the corresponding relations. The simplest form of graphical presentation is the *tree structure* (see example in ISO 5964). Tree structures can be produced manually, but some systems generate them automatically.

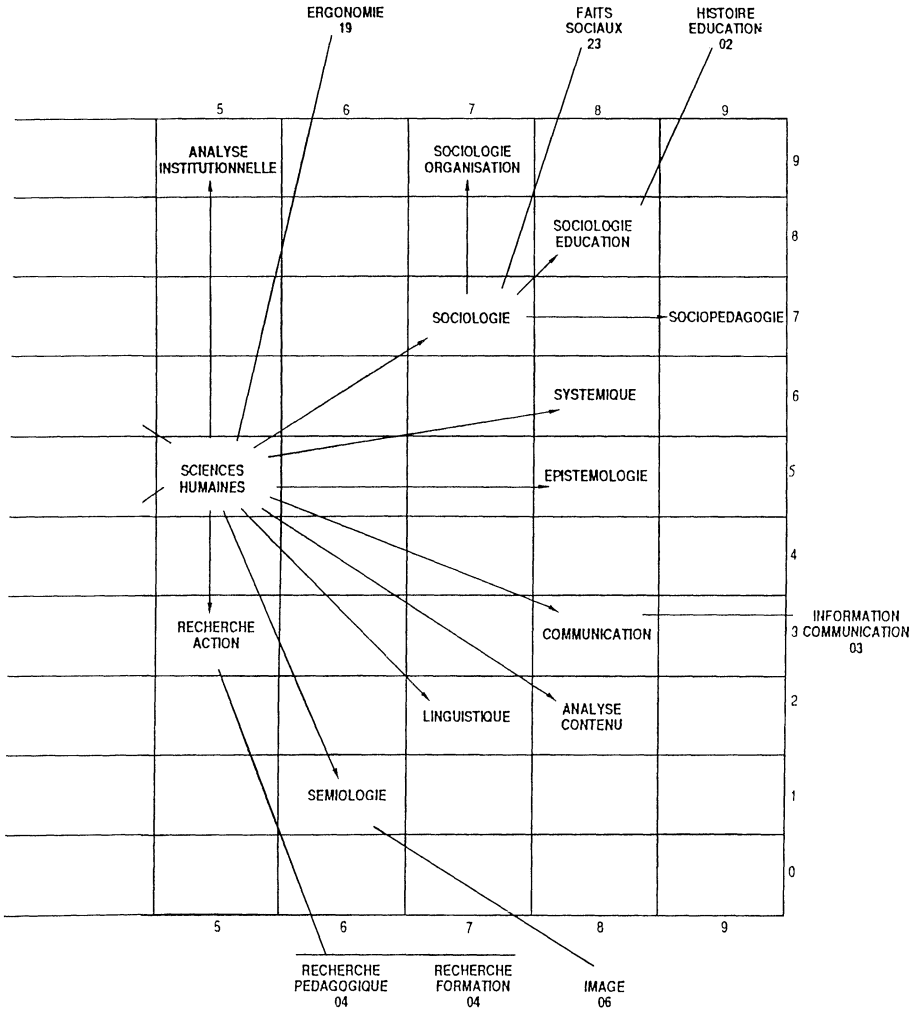
tree structures

A more sophisticated form is the *arrow graph* presentation (see 5), which is produced using computer-assisted design software. Textual data can be automatically loaded into such systems, but the actual position of the terms, arrows, and lines that illustrate relations has to be designed manually. In some thesauri, the coordinates of the descriptors in the "arrow graph" are given in other types of presentation. Consequently, it makes sense to integrate

arrow graphs

Figure 5: *Arrow Graph*




ERGONOMIE
19
FAITS
SOCIAUX
23
HISTOIRE
EDUCATION
02
5
6
7
8
9
ANALYSE
INSTITUTIONNELLE
SOCIOLOGIE
ORGANISATION
SOCIOLOGIE
EDUCATION
SOCIOLOGIE
SOCIOPEDAGOGIE
SYSTEMIQUE
SCIENCES
HUMAINES
EPISTEMOLOGIE
RECHERCHE
ACTION
COMMUNICATION
INFORMATION
COMMUNICATION
03
LINGUISTIQUE
ANALYSE
CONTENU
SEMIOLOGIE
9
8
7
6
5
4
3
2
1
0
RECHERCHE
PEDAGOGIQUE
04
RECHERCHE
FORMATION
04
IMAGE
06

graphics creation and manipulation capability with general thesaurus management functions. In practice, however, such data are encoded manually in the thesaurus system after the graphical display has been stabilized.

terminographs

Terminographs offer another form of graphical presentation (see Figure 6). The example given has actually been produced semi-automatically using a sophisticated, but proprietary system. Such a presentation can also be produced manually using a computer-assisted design package.

4.2.2 *Electronic Formats*

data interchange

Electronic formats are especially relevant in situations where the thesaurus is managed and used in different systems. Where both systems are built on the same database management system or can handle the same database file format, and if both are under total control of the same body, any "internal" format may be used. For external purposes, standardized formats, such as ISO 2709 or FORMEX, can be used for communicating data to targets systems. In such cases, the ability of addressees to correctly handle these very sophisticated formats can limit operational effectiveness. Partners may also agree on any specific format, in which case:

- either the thesaurus manager takes responsibility for providing a format that will allow straightforward loading of data into identified target systems ("specific target systems formats")
- or the thesaurus manager provides a single, highly readable, "flat" text file with "tags", together with the explanation of the tag names used, possibly introducing some data redundancy to facilitate easier data loading; such files assume no responsibility for further use of the data provided beyond the guarantee of completeness and internal consistency ("tagged text files").

4.3 THESAURUS MAINTENANCE

4.3.1 *Recording Modification Requests*

modification request forms

Thesaurus managers generally use special forms for collecting or recording user requests for modification. Any forms circulated among the user community must be designed to achieve the right balance between the point of view of the users (who want to be able to report easily any problem they encounter) and that of the

thesaurus manager (who looks at modifications in a more technical way and usually needs more complete data).

automating request management

Obviously, automating this process increases efficiency, and the number of proposals to be processed will determine to what extent such automation is useful. Automating the management of modification requests may range:

- from providing a specific encoding module, which "translates" requests as received from users into sets of data that can be handled by the system
- to automated pre-processing of envisaged modifications, which generates special reports with informative and warning messages
- to automation of clerical work, which features automatic production receipt acknowledgement and reports on the state of processing

4.3.2 *Recording Decisions*

modification history

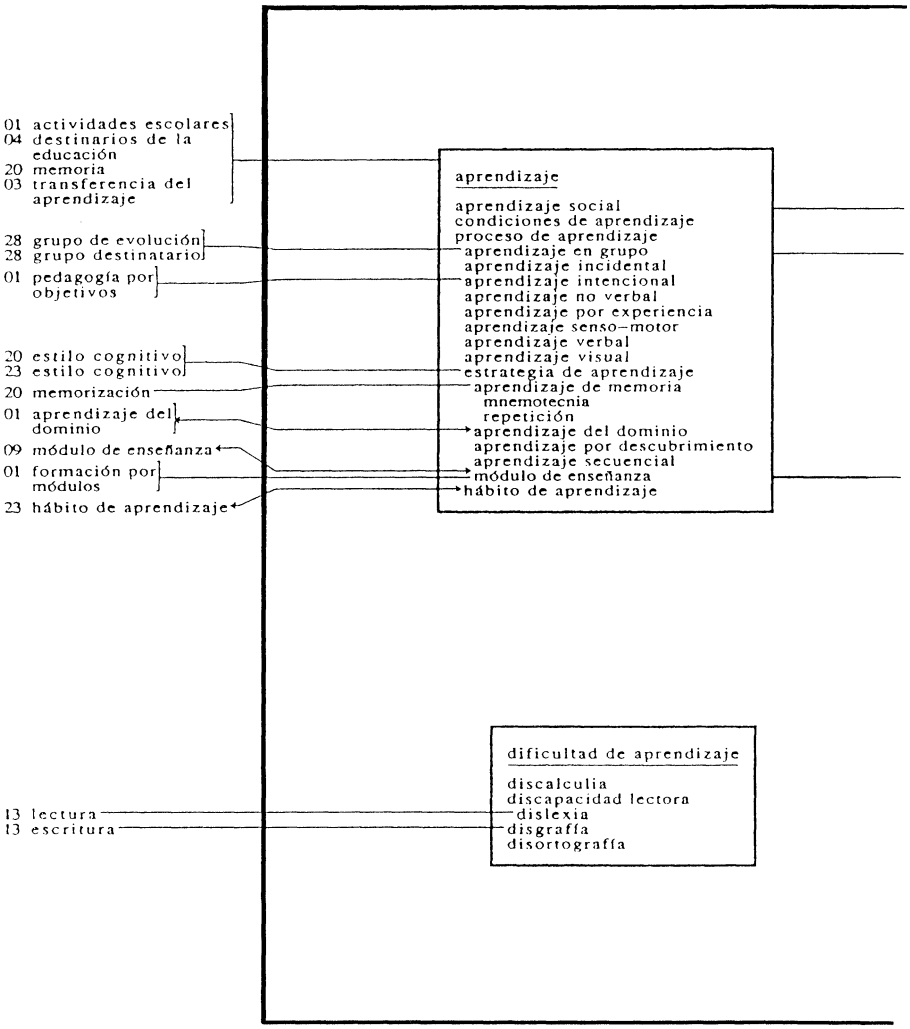
Decision records should document which modifications have to be taken into account by the system, but decisions on modifications include decisions *not* to modify the thesaurus, and these decisions should also be recorded because simple deletion of related data would prevent the manager from:

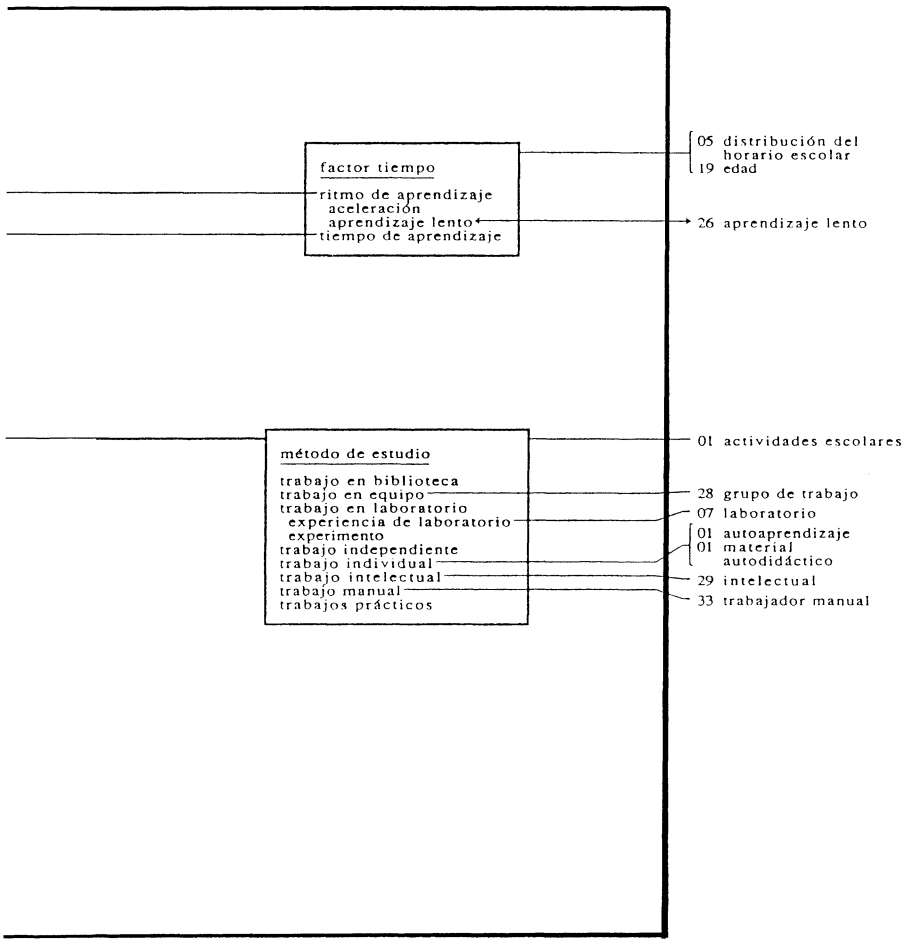
- informing proposers about the processing of their requests
- relying on previous decisions when taking new ones

content management strategies

This second aspect is especially important. The quality of the recorded information will determine to what extent the thesaurus management system will contribute to the long-term coherence of strategies for managing the thesaurus content. This information should not be limited to the decisions taken (which can easily be coded) but should also include their motivation (which implies recording textual data). Management efficiency can be improved by making information available such as: "this term has already been proposed as a descriptor but has been rejected for the following reason ...".

Figure 6: *Terminograph*




factor tiempo
ritmo de aprendizaje
aceleración
aprendizaje lento
tiempo de aprendizaje
05 distribución del horario escolar
19 edad
26 aprendizaje lento
método de estudio
trabajo en biblioteca
trabajo en equipo
trabajo en laboratorio
experiencia de laboratorio
experimento
trabajo independiente
trabajo individual
trabajo intelectual
trabajo manual
trabajos prácticos
01 actividades escolares
28 grupo de trabajo
07 laboratorio
01 autoaprendizaje
01 material autodidáctico
29 intelectual
33 trabajador manual

4.3.3 *Issuing Update Products*

Section 4.2 discussed the production of complete editions of a thesaurus for the users and target systems; this section examines update-process products that are necessary:

- for managers and management bodies to help them in the course of the update process ("update-process working documents")
- for users and target systems
  - to allow them to use updated data even if complete editions cannot be produced as often as managers want to provide users with new data (i.e., as often as managers stabilize intermediary updated versions of the thesaurus) ("update products")
  - to allow them to take necessary actions for migrating to a new version even if they are provided with complete updated editions ("migration products")

decision strategies

4.3.3.1 *Update-process Working Documents*. Such documents are prepared for the use of those responsible for decisions on content modifications and have to be designed to fit management needs. This means that data have to be presented according to two different types of distinctions:

- the distinction between the nature of the various modifications envisaged
- the distribution among domains, because managers need to have a comprehensive view of the context of the decisions to be taken and of the possible interactions of successive decisions in the same semantic field so that they can organize their work accordingly

Types of possible modifications include:

managing modifications

- adding and deleting concepts
- adding and deleting hierarchical, inclusive and associative relations
- creating and deleting non-descriptors and scope notes
- modifying the wording of existing descriptors, non-descriptors and scope notes

In the management of a multilingual thesaurus, it is useful to distinguish between the following kinds of modifications. Some changes that concern all language versions, such as the addition

modifying multilingual thesauri

and deletion of concepts and modifications to structural relations, have to be examined collectively. Other changes, typically modifications to terms and texts, apply to one language version at a time and can be examined by language-version managers only if the management system has been organized to accommodate such a distinction. However, this capability cannot be based on a purely technical definition, such as limiting it to specific types of elements to be modified.

language version interaction

Some modifications envisaged for only one language version, such as the modification of a descriptor or the introduction of a scope note, may nevertheless be of interest for, or have important consequences in, the other language versions because they tend to modify the intended concept. Thus, the manager must have the option to assign the appropriate status to each modification still to be examined or completed. In this case, it is also necessary to produce documents that combine data from all language versions, even if they do not necessarily appear together in any printed format. Such a document might consist of a "multilingual concept form" incorporating not only all language equivalents but also non-descriptors and scopes notes from all language versions.

update management data

Update-process working documents also need to incorporate data that are relevant solely for management purposes, such as:

- modification status (documents will incorporate previously approved, pending, and newly proposed modifications)
- definitions that aid the understanding of the proposed concept but that will not necessarily be kept in the thesaurus itself if the concept is actually introduced
- comments made by users, managers, or consulted experts
- sources for proposals, terms, texts, and comments

It is beyond the scope of this chapter to discuss possible ways of actually presenting these data.

summary information

Finally, managers need summary information on the consequences of the decisions they have taken:

- Good thesaurus management systems allow for some early feedback during management meetings or soon thereafter.
- If prepared for management purposes, the kinds of possible comprehensive update documents described below may also take into account modifications still pending.