# Domain analysis,
# an important part of thesaurus construction

**Marianne Lykke Nielsen**
The Royal School of Library and Information Science
Institute of Information Studies
E-mail: mln@db.dk

## Abstract

Design and construction of index languages should be based on thorough knowledge and understanding of the environment of the information system and its users. This study investigates a mixed set of methods (group interviews and word association to collect data, content analysis and discourse analysis to analyze data) to evaluate whether these methods together collect the data required for user-oriented thesaurus construction. Based on a case study we find that the three studies together provide the needed domain knowledge to define the role of the thesaurus and design the content. The analysis also provided conceptual knowledge and prepared for the next step, collecting and analyzing concepts and terms, thus saving resources in the construction process. In sum, the domain study made it possible to design a thesaurus reflecting the work environment and its users.

## 1   Introduction

The purpose of this paper is to demonstrate the usefulness of a multi-pronged domain analysis for establishing a firm basis for user-oriented thesaurus construction.  It has long been acknowledged that indexing practice and controlled index languages should be based on knowledge about the information system and its users (Mooers 1958; Soergel, 1974, Soergel, 1985; Aitchison, 1997; Bates, 1998; Jacob & Shaw, 1998). Several aspects must be considered:

- the system users and their information needs and search behavior,
- the subject field,
- the use of language,
- the type of data/literature, its quantity, and the available resources.

The literature does not give much advice on how to collect the requisite information about the environment. Soergel (1985) sketches some approaches to studying information needs as a basis for information system design. Hjørland (1997) operates with the concept of "domain analysis" on which system development and improvements should be based. Like Soergel, he primarily recommends what type of knowledge to gain about the information system and its users, but gives only few details about what investigative methods to use to collect the desired knowledge. Jacob and Shaw (1998) recommend the use of structured observation,

content analysis, work-flow study, and task analysis, but give no details about the exact use of the methods. Only very few research papers discuss the analysis needed at a methodological, practical level. The aim of this paper is to take up the thread and evaluate a mixed set of methods: group interviews and word association to collect data, content analysis and discourse analysis to analyze the collected data. The idea is to investigate whether these methods together collect the data required for user-oriented thesaurus construction, and how it is possible to implement the results in thesaurus design. The methods are tested in a real-life environment as part of a concrete thesaurus project in a large product-development company, working within the pharmaceutical industry.

The next section discusses the concept of domain analysis, Section 3 presents the case study and the study methodology (which was to be tested), Section 4 reports the study findings, and Section 5 illustrates the usefulness of this type of study for thesaurus construction. Section 6 summarizes the paper.

## 2. Domain analysis

Controlled index languages like thesauri have an important role in IR as devices which represent and provide access to documents and information. Both indexers and searchers are guided by the controlled vocabulary, which provides a map of the given knowledge field and guide the users so that they choose appropriate index or search terms. The major purpose of a thesaurus is to provide a map of a given field of knowledge, indicating how concepts and ideas about concepts are related to one another, and thus to help indexers and searchers to understand the structure of the field. A thesaurus represents the vocabulary of a specific subject domain, and it is important to develop the vocabulary according to the domain, the information system, and its users (Soergel, 1985; Aitchison, 1997, Jacob & Shaw, 1998).

The thesaurus design also depends on indexing policy and practice. The index language should fit indexing, and decisions concerning the two belong together and cannot be studied independently. In designing a thesaurus one must decide what indexing method to use, who is going to do the indexing, and how exhaustive the indexing should be (Jacob & Shaw, 1998). The focus of this paper is thesaurus construction; the connection to indexing practice will only be mentioned when necessary. However, an investigation of the information system should cover both aspects and lead to conclusions about both the index language and indexing practice.

## 2.1. Perspective: Person-in-situation

In the literature about thesaurus construction, this kind of investigation has no specific name, and the guides do not prescribe much about what methodology to use. Hjørland (1997) was among the first to introduce the concept of domain analysis as the basis for design and improvement of information systems. He works from an activity-theoretical approach, and for him the object of domain analysis is the development of collective information and knowledge structures. A typical domain analysis might examine the information structure of a discipline, including the size of its literature, the distribution of the literature with respect to various publication forms, its national/international structure, its citation patterns, disciplinary exchange etc. Hjørland's concept of domain analysis implies a criticism and skepticism towards more traditional user studies which he considers to be individualistic and defective, because they are often based solely on studies of individual users' information

needs and information-seeking behavior. In Hjørland's view, individual knowledge structures and worldviews are reflections of the objects of the work communities and their role in society, and individual information needs, mental models, search behavior, and relevance criteria should be analyzed in this perspective. Tom Wilson developed in 1981 a model of the information-seeking process, illustrating the factors which are influencing the IR process. The model illustrates very well that the searcher, trying to find information to satisfy a need, is bound by barriers of different kinds. Some barriers are personal, others are role-related, and others again are related to the specific environment in which the searcher is acting (Wilson, 1999).

Research into social and situational influences on information needs and user behavior has demonstrated that the situations in which people find themselves have a profound effect on their information-seeking behavior, but research also shows that high-knowledge users differed from low-knowledge users, or how users with different cognitive styles interact differently with information systems. Regardless of the situation of an individual, there are also individual variables that influence how that person acts (Furnas, 1987; Ingwersen, 1996; Iivonen, 1995; Allen, 1997; Bates, 1998, Ennis, Sutcliffe & Watkinson, 1999).

Personal and contextual variables interact during the search process, and the analysis of the environment of the thesaurus should incorporate the individual as an information seeker inside a particular environment and in a particular situation. The thesaurus is a tool that supports the individual user to get an understanding of the structure of the knowledge domain, and thus domain analysis should cover both the characteristics of the domain in order to present the view and perspective of the particular knowledge domain in the thesaurus, and the characteristics of the individuals who are acting in the domain community to be able to support them in the retrieval process. Individuals are shaped by their education, by former work experience and by other social and cultural experiences, and even though they try to adapt to the situation and the environment in which they are acting, they have been shaped individually and will approach the knowledge domain from an angle which depends both on their actual situation and on their individual characteristics. Allen (1997) calls it a person-in-situation approach, and in our view a domain analysis should be carried out from that perspective.

## 2.2. Aspects of domain analysis

In the present work, domain analysis is seen as the natural starting point of an iterative thesaurus construction process which consists of several interacting sub-processes. The study is based on the following model of the thesaurus construction process:

1 Determination of context, objectives, content, and design;

2 Analysis and choice of software;

3 Collection of concepts and terms to be analysed;

4 Analysis, choice, and structuring of concepts and terms;

5 Introduction and test of the thesaurus;

6 Maintenance

Two types of knowledge are needed for thesaurus construction: domain knowledge and conceptual knowledge. Domain knowledge covers knowledge about the environment and the situations in which the thesaurus is going to be used, knowledge about work tasks, knowledge about information use and information needs, and knowledge about the different approaches to the subject field. Domain knowledge is essential to defining and designing the thesaurus. Conceptual knowledge is linguistic, semantic knowledge about naming, form, meaning and relations between concepts and terms. Conceptual knowledge allows us to choose, control, and structure the vocabulary of the thesaurus.

The domain in which the thesaurus is going to work may vary, and the investigative design must be tailored according to the given situation. A domain may be embedded in cultural traditions, in a professional environment, or in a scientific environment (Ørom, 2000), and domain analysis must be shaped according to the domain characteristics. In this paper, a company constitutes the domain. This is a professional work domain, and the analysis focused on investigating the following factors:

- the nature of the professionals (background, work tasks, information needs, information use, language use, searching behavior[1], search problems),

- the subject field (topics, concepts, vocabulary),

- the literature (type, level, quantity), and

- the available resources for indexing and thesaurus construction (competence, time).

The main focus has been to get an understanding of the information-related work tasks[2] and the derived information needs. In the professional work domain, work-tasks provoke and have a strong influence on information use and retrieval; to design a tool that supports professional information searching activities we must consider the professionals' work tasks and, most importantly, their understanding and use of information in relation to their work tasks.

---

[1] Wilson (1999) has developed a model of the information seeking and information searching research areas. The model distinguish between *information behaviour*, defined as the more general field of investigation, *information-seeking behaviour* as a sub-set of the field, particularly concerned with the variety of methods people employ to discover, and gain access to information resources, and *information searching behaviour*, defined as a sub-set of information-seeking, particularly concerned with interactions between information user and computer-based information systems. Domain analysis focused on investigating the information searching behaviour as the thesaurus is going to support information retrieval in computer-based information systems.

[2] In this paper a work task is considered as the task either given to, or identified by, a worker. A task has a recognisable beginning and end, and it contains stimuli and guidelines concerning goals and/or measures to be taken. A task may have different levels of complexity and may consist of a set of subtasks (Byström, 1995). In the context of thesaurus construction, we are primarily interested in work tasks that involve information retrieval.

## 2.3. Methods

Several methods may be used to carry out a domain analysis, and many require a set of multiple methods to investigate a certain environment. A combination of multiple research methods will reveal a more varied and valid picture of the research object (Fidel, 1993; Vakkari, 1997). Qualitative methods are best for exploring human behavior, as the purpose of qualitative research is to describe how people behave and understand why they behave the way they do (Kvale, 1994). In the present study we carried out several analyses, qualitative as well as quantitative, for a complete domain analysis:

1. group interviews,
2. content analysis and discourse analysis of user requests, and
3. word association test.

In order to obtain an understanding of the work domain and its users, we interviewed 29 future users of the thesaurus. We analyzed 50 user requests using the methods of content analysis and discourse analysis to investigate from which perspective and aspects the users approach the particular subject field. Finally, we carried out a word association test in order to identify language use and approaches to the subject field.

## 2.4. Benefits of domain analysis

Taken together, the three studies gave a good picture of the work domain, but none of the tested research methods alone provides sufficient data for exploring and understanding the work environment.

Interviewing mostly provides domain knowledge. Work tasks may be described thoroughly providing insight into the context of information needs and search behavior. The actual discussions provided detailed knowledge about search routines and search problems.

The request analysis also provided good insight into the work domain; it showed that the work domain can be divided into 5 types of information-related work tasks, each characterized by a specific discourse and a set of typical information needs (see Figure 3). The interviews did not provide the same clear picture of the work tasks, but provided information to interpret and understand the observations of the discourse analysis and the content analysis of requests.

The content analysis of search requests has functioned as an essential part of the discourse analysis. The content analysis gave a picture of language use, but not of the meaning and understanding, and the combination of getting an overview by the content analysis and an understanding by the discourse analysis was very fruitful with respect to human and time resources. In sum, the request analysis has provided domain knowledge and insight which is useful primarily for designing the thesaurus and not for controlling the words of the thesaurus. However, the investigation has also resulted in conceptual knowledge about word form, concept categories, and relations to be used in the later semantic analysis.

The word association method provides some domain knowledge. It reveals domain-specific relations between words. But the results of the word association method must undergo

qualitative analysis to provide an understanding, and so word association provides material for subsequent qualitative analysis rather than final useful results. On its own, word association is a useful method to collect conceptual knowledge for the vocabulary and semantic control of thesaurus terms. Having identified the core concepts of the thesaurus by the content and discourse analysis, the word association method seems to be an economic way to collect synonym variations of the core concepts, and a way to find more domain-specific relations to the concepts, including local (slang) variations. The sample of thesaurus records (see Appendix D) includes several terms that have been identified by the pilot association test; for example, for *Adverse events* it includes the Related Terms *C-T prolongation* and *treatment withdrawal*. These related terms reflect specific interests of the work domain, and they will probably not appear in other thesauri within the subject field.

Before the domain analysis had been carried out, several sources had been consulted in vain to see whether they could form useful source material. Not until the domain analysis did the sources for term collection and analysis become clear.

In sum, the domain analysis has provided the needed knowledge about the environment of the thesaurus and made it possible to define the role and content of the thesaurus. The analysis has supported the difficult task of defining the subject field (focus and approaches), and of deciding what types of information and relations to include (content and structure).The findings clearly demonstrate what subject areas to focus on, and the domain analysis prepared for the next step, collecting and analyzing concepts and terms to be included in the thesaurus, thus saving resources in the following construction process. By the analysis it became clear that the local discourse is formed by terms from a few concept categories, and these findings have helped the thesaurus manager to define a focus. Selecting the set of associatively related concepts is normally very difficult; here the analysis of requests has revealed some patterns between certain concept categories that may support the structuring so that it reflects the work domain. Furthermore, the association test has proved to be a useful method to identify related terms.

The next section describes the case study and the methods in detail, giving for each method *the purpose of the analysis, materials and methods of data collection*, and *methods of analysis*. This is followed by a section on the study findings and their implications for thesaurus design.

## 3. Case study

The empirical study reported in this paper is part of a thesaurus project which aims at developing a corporate thesaurus for a company that develops and produces pharmaceuticals for psychiatric and neurological disorders and for the treatment of diseases of the central nervous system (CNS) to support information storage and retrieval in a document management system. The objectives of the system are as follows: facilitating the exchange of documents and information across the company, thus increasing the exchange and use of knowledge across different departments and work tasks, support the production electronic medical submissions to legal authorities, and generally support and facilitate information searching in internal and external information systems.

Different types of documents will be stored in the future document system: Standard Operating Procedures (SOPs), research reports, test reports, articles, statistics, etc. The documents come from different departments and they are produced by people who differ in their approach to the common subject field, specialist knowledge, work tasks, and language use. Each type of document requires a specific set of metadata.

The documents will be indexed by a mix of novice and expert indexers, and, thus, it will be difficult to produce document representations of consistent quality. Also the searchers differ in search competence and behavior. Having these constraints in mind, we decided that the thesaurus shall support free-text retrieval as well as retrieval based on manual, controlled indexing.

## 3.1. Group interviews

### 3.1.1. Purpose

The purpose of the interviews was to gain an overall insight and understanding of the information-seeking behavior in the work domain. The study focused on getting information about information needs, information sources, search behavior in electronic information sources, and problems in relation to the electronic search process. We collected oral descriptions and group discussions of information use and search.

### 3.1.2. Methodology

We used the methodology of the qualitative research interview, specifically open group interviews. The group interview was based on an interview guide, structured into 6 categories of questions:

- Participants: educational background, position, work tasks, information use
- Information needs
- Information resources
- Search behavior
- Search problems
- Search aids

We carried out 8 group interviews involving 29 participants; each lasted two hours. The participants were chosen according to the following criteria (purposive sample):

- Electronic information searching is a natural part of their work tasks.
- They have experience and knowledge about the problem area.
- They are a good representatives for their work task domain.

The group interviews were carried out in such a way that every participant had the opportunity to answer all the questions individually. The main questions were planned and structured before the interviews, but the process of answering the questions was conducted as a free process so that the participants had the opportunity to develop the questions during the conversation. If a new question arose from the conversation, every informant was asked to

comment on the problem. The interviews were taped. To be sure that every informant had the opportunity to come up with his specific problems and angle, they were asked to write down an answer in a written questionnaire, containing open as well as closed questions. This was done in order to avoid a single participant dominating the interview.

### 3.1.3. Analysis

Shortly after each session we prepared a short interview description – problems, discussions, and viewpoints – based on the tape recording. We analyzed the questionnaire responses and developed some statistics from the closed answers. The project group as a whole analyzed the interview material and summarized them in a report, which was later presented to a set of participants and discussed with them in order to clear up misunderstandings (Lykke Nielsen, 1999).

### 3.2. Analysis of user requests

### 3.2.1. Purpose
The purpose of the request analysis was twofold:

a)  to obtain knowledge about the work domain in order to plan and design the thesaurus, and

b)  to gather terminological and semantic knowledge for the later conceptual analysis.

### 3.2.2. Methodology: Overview

The data collection consists of a randomly selected sample of 50 user requests received by the corporate librarians in a two-year period from August 1998 – May 2000. The requests were sent by e-mail to the library, which is responsible for obtaining material from external resources. The request statements (Figure 1) are often very short and written in a personal and confident style, but they are considered as adequate descriptions regarding context and form of the corporate

---

**From:** Michael Jensen
**To:** Pia Jørgensen
**Date:** 12. Januar 2000
**Subject:** Phanquinone

Dear Pia,
I am looking for information about a substance, which is called Phanquinone. It is not a new substance, but it has been known for some years. I want to know something about the analysis of it in serum and plasma, and I want to know something about its pharmacokinetics in animals and/or in humans.
I suggest the following keywords: phanquinone, pharmacokinetics, quantitative determination of (in plasma/serum), HPLC.

Best regards
Michael

---

**Figure. 1: Example of e-mail request**

information needs, because long collaboration has led to a common understanding of the context of the underlying work tasks and related information needs between the researchers and technicians on the one hand and the librarians on the other.

The data were analyzed using two methodologies: quantitative content analysis and qualitative discourse analysis interpreting the context of the request.

### 3.2.3. Purpose of the quantitative content analysis

There are two approaches to content analysis in library and information science. One approach, *classification analysis*, classifies documents according to their content using either a pre-existing classification scheme or a novel one. The second approach, elemental analysis, is based on the identification of word or word group frequencies (Allen & Reser, 1990). We conducted an elemental analysis to obtain quantitative knowledge about:

- Types or categories of concepts in the requests
- Number of concepts and concepts types
- The relations between concepts
- Terminology

### 3.2.4. Methodology for the quantitative content analysis

The content analysis was carried out using the following procedure. First, the requests were divided into elemental concepts by a facet analysis. Then each elemental concept was categorized according to a list of categories based on the MeSH Tree Structure. To give an example, the request of Figure 1 was divided into 5 elemental concepts categories or facets (see Table 1).

**Table 1**
Facet analysis of requests

| Elemental concepts | Concept types |
| --- | --- |
| 1) phanquinone | C. Chemical and drugs |
| 2) pharmacokinetics | D. Pharmacology |
| 3) in animals/humans | A. Organism and test system |
| 4) quantitative determination in plasma/serum | J. Analytical methods and techniques |
| 5) HPLC | J. Analytical methods and techniques |

In order to investigate the relationship between concept categories and to see if some combination patterns exist, we constructed a request-category matrix (Appendix A) and calculated the frequency of each category (Appendix B). Finally, we compared the vocabulary across the requests and to the vocabulary of the MeSH thesaurus to investigate language use. We analyzed the choice of terms to see if the requesters use scientific terms, slang words, abbreviations, etc.

### 3.2.5. Purpose of the qualitative discourse analysis

The objective of the qualitative analysis was to investigate the users' way of understanding and relating the concepts of the work domain in order to determine whether the requests represent specific perspectives or conceptions of the subject field. The qualitative analysis focused on the context of the needs:

- the requester's educational background,
- related work tasks,
- use of information,
- importance of information,
- approach to subject field.

Users presumably approach the subject domain from different angles depending on their educational and former work experience and on their present work tasks. At the same time, the individual user probably approaches the topics from a perspective which he shares with other persons working within the same work task context. Influenced by their common work tasks, users may form a discourse community and approach the subject field with a common understanding. Determining whether discourse communities exist within the work domain is the main objective of the qualitative analysis. As a complement, we want to find out whether two levels exist: an individual level connected to individual educational background and work experience, and a collective level connected to specific work tasks or knowledge domains. The qualitative analysis should answer the following questions:

- What concepts are important within the work domain?
- From which perspective do users approach the concepts?
- Why do users relate the concepts?
- Are any concepts or aspects more important that others?
- Do users have a common understanding of the concepts?
- Do users have a common terminology?

The answers should make it possible to decide whether different discourses or sublanguages occur in the work domain and to decide whether the differences are so strong that the thesaurus should reflect, articulate, and bridge the differences.

### 3.2.6. Methodology for the qualitative discourse analysis

The analysis of the requests was inspired by the theory and practice of discourse analysis. A discourse is defined as a way to talk about and understand experiences and concepts within the social world. Often different discourses fight against each other with the wish to freeze one particular meaning and understanding of the world. In traditional discourse analysis, the contrasts and conflicts of different discourses are the main study objects, including the interactive and social processes which create the different discourses and conflicts (Winther, Jørgensen & Philips, 1999).

Discourse analysis has recently been introduced in library and information science: Budd and Raber (1996) have used it to investigate the social, political, and technical uses of the word "information" to study the implications for theory and practice, and Talja and her colleges (1997) have used it to build the search vocabulary of a www-based local information service by extracting from client-intermediary dialogues the search terms and dimensions by which users approach the information service.

Our data for the qualitative analysis – written search requests – are not the typical material for a discourse analysis. Several kinds of material are commonly recommended as the basis for discourse analysis: "naturally available materials" and "collected materials" (Potter, 1999). Naturally available materials are research texts, media texts, and transcriptions of everyday conversations; these kinds of materials are considered to be neutral since the researcher does not have any opportunity to influence the statements of the text. Collected materials are the product of interactively produced statements obtained in semi-structured or unstructured interviews.

We use written requests as basis for our discourse analysis (naturally available material). However, the requests are generally very short and do not provide much information to do a proper discourse analysis. So in order to establish more context for our interpretation, we used other sources to obtain data on the characteristics of the requester. The sources consulted include: internal research reports, guidelines, encyclopedias, and often interviews with the librarians who have handled the requests.

A typical analysis started with identifying the requester, his organizational position and work tasks. We consulted the corporate address book and interviewed the librarian who had taken care of the request in order to get background information about the underlying information need and the related work task. We consulted related guidelines, reports and SOPs (Standard Operation Procedures) in order to study the definition and understanding of the concepts and to investigate the importance and use of the concepts in the particular work context. In this way we were able to put the requests and concepts into a context and understand the viewpoints and conceptions on which the use of the vocabulary was founded.

### 3.3. Word association method

### 3.3.1. Purpose

In the present project, the aim is primarily to use the method for the later semantic analysis, but a pilot test was carried out to test whether it may be fruitful also to use the method as part of domain analysis to investigate whether domain-specific approaches to the subject field exist within the work domain.

The word association method is a quantitative method that originated in cognitive psychology and linguistics. It has been used only sparingly in information science (Reisner, 1966; Rubinoff, 1967; Pejtersen, 1989; Lykke & Skrubbeltrang 1992). In its simplest form a series of disconnected words (stimulus words) are presented orally or in writing to the respondents, who must respond with the first word that comes to mind (response words). These associations reveal the respondents' mental models and verbal memories, and provide the researcher with information about the respondent's way of relating terms and their use. The word association

method is not recommended as a suitable method to build up a detailed mental map. Aitchison (1994) mentions a number of problems, but the most serious shortcoming of word associations is that they cannot tell us about the probable structure of the human word-web, partly because each person is asked for only a few responses to a particular stimulus, and partly because the links between words are multifarious. One can never be sure to get all relevant relationships to a certain concept. Regarding the multifarious possibilities of linking, experiments show that it is possible to prime the respondents so that the associations are made according to a specific context (Aitchison, 1994; Lykke Nielsen & Ingwersen, 1999). Normally, the response words of a controlled, primed test will be based on the respondent's work experiences and on the daily language and reflect the work domain.

### 3.3.2. Methodology

We conducted a pilot test with three researchers as respondents. 24 stimulus words were tested; they were selected according to the findings about important concept categories. The stimulus words were presented both written and spoken. The respondents had 1 minute to give two word associations. They were instructed to make associations according to their work tasks and to concentrate on words related to the stimulus word from a practical perspective of use and application rather than from a formal, scientific viewpoint. Furthermore, they were instructed to consider the response words as words which may replace the stimulus words in a search query.

| Stimulus word: Electrocardiograms | | |
|---|---|---|
| Response words | Type of relationship | Viewpoint |
| C-T prolongation | RT | Work domain |
| Sertindole | RT | Work domain |
| Side effects | RT | Pharmaceutical viewpoint |
| ECG | Synonym | Medical viewpoint |
| Heart rhythm | Synonym[3] | Medical viewpoint |
| Heart function | RT | Medical viewpoint |
| Safety pharmacology | BT | Pharmaceutical viewpoint |

**Figure 2: Response words to the stimulus *Electrocardiograms***

---

[3] In this particular context the word *heart rhythm* is considered to be synonymous with *electrocardiograms*

### 3.3.3. Analysis

The analysis has focused on investigating the quality and the character of the associations. The quality was evaluated by respondents themselves, who immediately after the pilot test were asked to judge the quality and relevance of the associations in relation to the work domain. The 24 stimulus words elicited 153 unique response words; only 6 of these were judged to be irrelevant to the context of the work domain. Afterwards the response words were divided into types of relationships; as may be seen from Figure 2, the response words consisted of a mix of synonyms and associatively related words, among the words categorized as synonyms some generic broader and narrower words appear.

In order to check if the word association method reveals domain-specific relations between the stimulus words and the response words, the response words were divided into categories of scientific viewpoints: medical, pharmaceutical, and domain-specific.

## 4. Findings

The following sections present the findings according to the aspects addressed by the domain analysis. The presentation will focus on the users, their work tasks and information needs; language use; and search behavior and search problems. Section 5 discusses the implications of the findings for thesaurus design.

### 4.1. Work tasks and information needs

The users fall into 5 types of information-related work tasks (see Figure 3), each representing an approach to the overall subject field of the work domain. Each specific work task group has its own discourse and some common, typical information needs.

A work task group is formed by professionals, possibly from several departments in the organization, who share the same overall mission and duties. The shared mission influences the way the members of the group search for and use information. The members of a work task group approach the overall subject domain from the same angle; they constitute a discourse community[4]. The members of a work task group have similar information needs. The work tasks and the derived information needs can be described according to the following characteristics: a priori determinability, repetitiveness, novelty of concepts, and type of information need (Byström & Järvelin 1995). *A priori determinability* refers to the degree of a priori uncertainty about the task, process, and outcome. A low degree of a priori determinability indicates that neither the result, the process, nor the information requirement can be characterized in advance. A high degree of *repetitiveness* indicates that the task involves actions that are repeated many

---

[4] A discourse represent a specific way of talking about and understanding the world (Winther & Jørgensen, 1999). A discourse community, is a group of people who, at least in the context of a particular role, hold a recognized body of "truth statements" in common (Olsson, 1998).

| Work task type/discourse community: | Characteristics of work tasks and derived information needs: | | | | |
|---|---|---|---|---|---|
| | A priori determinability | Repetitiveness | Novelty of concepts | Common concept categories | Type and structure of information need |
| **All discourses** | | | | Chemicals & drugs | |
| **Basic research:** Development and test of new substances and chemicals -> composition & effect | Low degree of determinability | Low repetitiveness | Known and unknown concepts | Pharmacology Persons External relations | Conscious information needs Few interrelated facets |
| **Clinical and non-clinical tests:** Test of new drugs in order to document effect and safety -> reaction | Low to medium degree of determinability | Some repetitiveness | Known concepts | Analytical methods Test systems Pharmacology | Conscious information needs Many interrelated facets |
| **Marketing and sales:** Sell and promote the effect and value of a product -> benefit | Medium degree of determinability | High repetitiveness | Known concepts | AE & reactions Diseases Pharmacology | Conscious information needs Some interrelated facets |
| **Competitive intelligence:** Compare the effect and value of a product with other products ->effect & benefit | Medium degree of determinability | Some repetitiveness | Known concepts and unknown concepts | Diseases External relations AE & reactions | Conscious information needs Some interrelated facets |
| **Environment and safety:** Secure the safety of the employees -> safety | High degree of determinability | High repetitiveness | Known concepts | Working conditions | Conscious and verificative information needs Few interrelated facets |

**Figure 3: Types of work tasks and information needs**

times and thus may be characterized as routine tasks. The tasks are case-dependent, but the process and the information requirements are well-known. *Novelty of concepts* refers to the degree of prior knowledge about the concepts involved. Prior knowledge is vital in determining what information is needed to accomplish a task (Bystrøm & Järvelin, 1995; Vakkari, 1999). This implies that the degree of concept knowledge is a major factor that determines the searching behavior and the need of conceptual support. If a user has insufficient knowledge about a task, he does not know the concepts and relations between concepts that are needed to formulate a query. Tasks of low determinability, low repetitiveness and high novelty require more conceptual support than tasks of lower uncertainty, and thus it is important to identify the characteristics of the works tasks and the derived information needs in order to design a thesaurus. The information needs have also been divided into 1) verificative needs, 2) conscious, topical needs, and 3) "muddled" information needs (Ingwersen 1982).

The works tasks connected to *Basic research* are largely indeterminable, not repetitive, and neither the process nor the information requirement are known in advance. The basic researchers develop new products or processes and thus investigate new and unknown topics. However, they do not consider their information needs to be "muddled" or the concepts involved to be unknown; they know very well what they are looking for. They prefer to make very broad searches in order to review a large set of retrieved literature; their requests/queries often consist of only a single chemical substance combined with the name of a researcher to see if the research of a known colleague or research institution may provide some help (see

Table 2). Similar results have been reported by other studies of R&D scientists (Hertzum & Pejtersen, 2000).

**Table 2**
Request formulations

| Work task group | Sample citation |
| --- | --- |
| Basic research | "I want information about J. E. Maggio, Harvard Medical School, Boston. He has written about Beta-amyloid". |
| Clinical and non-Clinical testing | "...a search on chlorprothixene, pharmacokinetics, plasma/serum concentrations (in patients)" |
| Marketing and sales | " literature on the treatment with Clopixol and/or Fluanxol to diabetic patients" |
| Competitive intelligence | "a search on Citalopram, Sertraline, Paroxetine, Fluxetine, Fluvoxamine, Venlafaxine and panic disorder" |
| Environment and safety | "a search on CAS 55501-05-8. I am interested in name and structure" |

The requests of the test scientists *(Clinical and non-clinical testing)* have a much more complex structure and consist of words belonging to several concept categories, but their work tasks are more a priori determinable and are more repetitive. Their information requirements vary with respect to novelty, but may still be characterized to be conscious topical needs, representing users who want to clarify or review aspects of a known subject matter. In general they deal with the reactions to drugs, and they require information that covers several interconnected facets:

1. the influence of a chemical or drug *(pharmacology)*,
2. under certain conditions *(diseases, persons)*'
3. in relation to certain test objects *(test systems*: animals, tissues, persons),
4. verified by a certain study technique *(analytical methods)*.

The discourses of *Marketing and sales* and *Competitive intelligence* are very similar to the one of the test environment, but they are even more determinable and repetitive. Their work tasks are concerned with the benefit of a drug, and they look for information which evaluate and demonstrate the effect and value of a certain drug type and, especially, to assure that no serious adverse effects are connected to the treatment. They have to document the safety of different kinds of patients: pregnant women, diabetic patients, overweight patients, etc., demonstrated by legal, approved study reports as well as post-marketing reports on registered adverse reactions (Periodic Safety Update Reports). Their needs are clear examples of conscious topical needs involving the following facets:.

1. the effect and value of a drug (*pharmacology, adverse events and reactions*)
2. under certain conditions (*diseases, persons*)
3. verified by a certain study techniques (*analytical methods, document types*)

The discourse of *Environment and safety* concerning the local staff is very simple, and often the needs are expressed as verificative needs using the CAS no. combined with the concept of safety. The purpose is to assure the safety of working with a certain chemical.

## 4.2. Language use

The pharmaceutical industry is very regulated with a specific set of rules and guidelines for documentation and with a domain-specific vocabulary. However, the common, regulated vocabulary is not accessible by one single source, and a variety of vocabularies and guidelines from different source organizations must be consulted in order to get an overview of the terminology. As a consequence, the scientists use many synonyms for the same concept, and the same word is defined and used differently by the different work task groups. In practice, it is a difficult process to become familiar with the specific way of working and the regulated language use, and the individual user often has a limited knowledge of the standardized vocabulary. Language knowledge and use depend on the users' position and work tasks and on the seniority in the organization. The *basic research scientists* and the *test scientists* tend to have less vocabulary problems than *people from competitive intelligence and marketing and sales,* apparently due to training and routine in using the specific document structure and vocabulary. They all have the same educational background, but *people from competitive intelligence and marketing and sales* are not obliged to use the language in the same precise, standardized way in their daily work tasks, and thus they do not possess the same knowledge about the domain-specific language as the researchers. Many of the users see the thesaurus as a means of compiling and presenting the regulated, standardized vocabulary to new and less-experienced users.

The local naming of drugs creates a special situation as the name of a drug shifts several times during its lifetime. When a drug is under development it appears to have a set of numbers, each representing a specific composition. Later it gets a generic name, and it ends up having various trade names depending on the market, the solution, and the weight. A set of competing, rival names are also connected to the drug. The different naming causes some problems in retrieval, because the drugs were indexed differently in local as well as external databases, and because the researchers seldom have a clear picture of the life story of a drug. It is possible to find the same pattern for other concepts, e.g., analytical methods.

## 4.3. Search behavior and search problems

All our users say that they have a clear picture of their problem situation and the derived information need. The users know what they are looking for, and they seem to have no difficulties stating the information need. However, a majority of the users say that they often find it difficult to delimit their search problems when they are formulating the queries. Demands from the regulating authorities make it important to retrieve as much relevant literature as possible, and thus they try to capture new ideas and directions to follow during the search process. Their descriptions of the search process are in line with the IR models which emphasizes the changing

nature of information needs during the search process (Ingwersen, 1996; Bates, 1998; Vakkari, 1999).

Our typical searcher has some search experience and is able to use advanced search techniques, like combination of different search criteria and expansion by synonyms and spelling variants. They also posses knowledge about search tools like thesauri. However, they tend to employ the corporate librarians for computerized searching, a pattern which is known from other studies on search behavior (Ellis & Haugan, 1997).

## 5. Implications for thesaurus design

The concept category *chemicals and drugs* form the center of the five discourses. The research scientists deal with the composition and the effect of drugs *(pharmacology)*, the test scientists, the marketing and the sales people with the benefit and reaction to drugs *(adverse reactions)* in relation to specific *diseases*, proved by some *analytical methods* using specific *test systems*. The environment people deal with the safety of *chemicals*. These few concept categories form the local discourses, and they must be covered thoroughly in the thesaurus, and important relations between, for instance, a certain drug and serious adverse reactions should be shown. To support the information search activities effectively it is necessary to consider not only traditional thesaurus information like reference to broader, narrower and related terms. The findings show that the users need a lexicon more than a traditional thesaurus. The thesaurus should provide a map of the field of CNS (Central Nervous System) research in order to support the users to understand the structure of the field. Appendix D gives some examples of thesaurus records from the thesaurus which is under development. The general structure of a thesaurus record is shown in Appendix C.

For the vital concept categories (*chemicals and drugs, diseases, pharmacology, adverse events and reactions, and test systems*) the users need clear definitions according to the regulated, pharmaceutical vocabulary. Also the related source and quality system are included in the thesaurus records to explain the definitions and the terminological choices (see Appendix C and D).

The scope notes defining the concepts *test systems, drug substances* and *adverse events* are clear examples of definitions which reflect and explain the use of these and related concepts within this specific domain. In related thesauri, e.g., MeSH, the concept of *test systems* does not exist, and in MeSH the concept of *laboratory animals* is hierarchical related to *animals* and not to a concept related to testing (like *test systems*).

The work tasks and information needs of competitive intelligence and the marketing people tend to be very homogeneous, and thus it should be possible to support this user group by relating drugs, diseases, and adverse events according to the latest test results, as it is shown in the record of *Parkinson disease.* The information needs of basic research scientists and test scientists are more varied and concern subjects of a higher degree of novelty, and thus it may be difficult to foresee their needs and support them directly. The basic research scientists might profit from references to experts who have a strong relation to a certain concept (see

the record of *Lu 10-171*), and the test scientists can be supported by continuously updating the concepts covering test systems and standard procedures. The concept of *drug substances*, for instance, is a well-known concept, but the naming as well as the definition and relationship to other concepts have been under discussion. The different discourse communities within the work domain use the terminology differently. The records represent the official corporate viewpoint, which is based on the International Conference on Harmonization (ICH).

The changing names of drugs pose a special problem; it is very important to make clear to the users of the thesaurus that a drug may be searched under a large set of words which in one context may be considered as synonyms and in another context represent different meanings and uses. The record of *Lu 10-171* shows that the drug has a generic name *Citalopram* and several trade names, e.g., *Celexa* ™. The Lu number is the preferred indexing term, but often documents must been found by the alternative names. The range of narrower Lu numbers to *Lu 00-010* indicate the existence of a set of variants of more or less the same substance.

The searchers access a variety of internal and external information systems and services, and, thus need support to formulate and expand their searches in all types of systems, local as well as external, and the searchers need reference to a huge set of search terms to be used for query term expansion: synonyms and near-synonyms, trade names, broader terms, narrower terms, and related terms which indicate new perspectives and directions, see , for instance, the record of *adverse events*. Due to the terminological inconsistency, the variety of synonyms also has an important function as entry words to the thesaurus.

## 6. Concluding remarks

The objective of the present work was to evaluate a set of different methods for domain analysis and to discuss the possibilities of integrating them into the methodology of thesaurus construction. The development of a thesaurus should be based on a thorough understanding of the environment in which the thesaurus is going to serve. We tested two qualitative and two quantitative methods. The qualitative methods produce domain knowledge, which supports decisions regarding the overall design of the thesaurus, while the quantitative methods mostly provide conceptual knowledge, which supports the later semantic and terminological analysis. Moreover, qualitative and quantitative methods complement each other; their combination functions well in practice. As a case in point, the quantitative content analysis of requests provides a good overview, and it may be fruitful to begin a domain analysis by categorizing and counting the words represented in important texts of the domain and use the resulting structure to inform the design of qualitative analyses. We conducted the domain study from a person-in-situation perspective in order to develop a thesaurus which at the same time reflects the characteristics and perspective of the particular knowledge domain and supports the individual information searcher. We believe that the findings fairly well reflect the work domain and clarified the conceptual problems. The user must interact actively with the knowledge structure provided in the thesaurus to bridges the gab between the work domain and his own individual perceptions, tasks, and information needs.

## Acknowledgements

## References

Aitchison, J, Gilchrist, A & Bawden, D (1997). Thesaurus construction and use: a practical manual. London : Aslib.

Aitchison, J (1994). Words in the mind: an introduction to the mental lexicon. Oxford : Blackwell.

Allen, B (1997). Information needs : a person-in-situation approach. In: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 14-16 August, 1996, Tampere, Finland. London : Taylor Graham. 111 – 122.

Allen, B & Reser, D (1990). Content analysis in library and information science research. In: Library and Information Science research 12. 251 – 262.

Bates, M J (1998). Indexing and access for digital libraries and the Internet: human, database, and domain factors. In: Journal of the American Society for Information Science, 49 (13). 1185 – 1205.

Budd, J M & Raber, D (1996). Discourse analysis: method and application in the study of information. In: Information Processing & Management, 32 (2). 217 – 226.

Byström, K & Järvelin, K (1995). Task complexity affects information seeking and use. In: Information Processing & Management, 31 (2). 191 – 213.

Ellis, D & Haugan, M (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. In: Journal of Documentation, 53 (4). 384 – 403.

Ennis, A, Sutcliffe, A G & Watkinson, S J (1999). Towards a predictive model of information seeking: empirical studies of end-user searching. In: Wilson, T D & Allen, D K (ed.). Exploring the contexts of information behaviour. London : Taylor Graham. 96 – 115.

Fidel, R (1993). Qualitative methods in information retrieval research. In: Library and Information Science Research, 15. 219 – 247.

Furnas, G W, Landauer, T K, Gomez, L M & Dumais, S T (1987). The vocabulary problem in human-system communications. In: Communications of the ACM, 30 (11). 964 - 971.

Hjørland, B (1997). Information seeking and subject representation. London : Greenwood Press. 213 p.

Iivonen, M (1995). Consistency in the selection of search concepts and search terms. In: Information Processing & Management, 31 (2). 173 - 190.

Ingwersen, P (1982). Cognitive analysis and the role of the intermediary in information retrieval. In: R. Davies (ed.). Intelligent Information systems. Chichester: West Sussex: Horwood.. 206-237.

Ingwersen, P (1996). Cognitive perspectives of information retrieval interaction : elements of a cognitive IR theory. In: Journal of Documentation, 52 (1). 3 – 50.

Jacob, EK & Shaw, D (1998). Sociocognitive perspectives on representation. In: Annual Review of Information Science and technology (ARIST), 33. 131 – 185.

Kvale, S (1994). InterView : en introduktion til det kvalitative forskningsinterview. København : Hans Reitzels Forlag. 318 p.

Lykke, M and Skrubbeltrang, C (1992). Indekseringsstrategi og søgethesaurus for UMI-informationssystemet (Development of an indexing strategy and a searching thesaurus for the UMI systems). Aalborg: Royal School of Librarianship. 91 p.

Lykke Nielsen, M (1998). Future thesauri: What kind of conceptual knowledge do searchers need? In: Proceedings of the Fifth International ISKO Conference, 25-29 August 1998, Lille, France. 153 – 160.

Lykke Nielsen, M (1999). Undersøgelse af informationsbehov og søgeadfærd ved Lundbeck A/S (Intern rapport). 7 p.

Lykke Nielsen, M & Ingwersen, P (1999). The word association methodology – a gateway to work-task based retrieval. In: Final MIRA ESPRIT Working Group Conference, April. Electronic Workshops in Computing. Full-text available from WWW at: http://www.ewic.org.uk/ewic/workshop/view.cfm/MIRA-99.

Mooers, C & Brenner, C W (1958). A case story of a Zatacoding information retrieval system. In: R S Casey, J W Perry, M M Berry & A Kent (eds). Punched cards, their application to science and history. New York : Reinhold. 346 - 352

Olsson, M (1998). Discourse – a new theoretical framework for examining information behaviour in its social context. In: Proceedings of the Second International Conference on Research in Information Needs, Seeking and Use in Different Contexts. 13/15 August 1998. Sheffield, UK. 136 – 149.

Pejtersen, A Mark (1991). Interfaces based on associative semantics for browsing in information retrieval. Roskilde : Risø Laboratory.

Potter, J (1999). Discourse analysis as a way of analysing naturally occurring talk. In: Qualitative Research : volume II. London : Sage. 323 –341.

Reisner, P (1966). Evaluation of a "growing" thesaurus. Yorktown Heights : IBM Watson Research Center. 19 p.

Rubinoff, M (1966). A rapid procedure for launching a microthesaurus. In: IEEE, 9 (1). 8 - 14

Soergel, D (1974). Indexing languages and thesauri: construction and maintenance. Los Angeles, CA: Melville. 632 p. (Wiley Information Science Series)

Soergel, D (1985). Organizing information. San Diego : Academic Press. 450 p.

Talja, S, Heinisuo, R, Luukkainen, S & Järvelin, K (1997). Discourse analysis in the development of a regional information service. In: Proceedings of the 2nd British-Nordic Conference on Library and Information Studies, Queen Margaret College, Edinburgh, 1997. London: Taylor Graham. 109 – 128.

Vakkari, P (1997). Information seeking in context : a challenging metatheory. In: Proceedings of an International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 14-16 August, 1996, Tampere, Finland. London: Taylor Graham. 451 – 464.

Vakkari, P (1999). Task complexity, problem structure and information actions. Integrating studies of information seeking and retrieval. In: Information Processing and Management, 35 (6). 819 – 838.

Wilson, TD (1999). Models in information behaviour research. In: Journal of Documentation, 55 (3). 249 – 270.

Winther Jørgensen, M & Philips, L (1999). Diskursanalyse som teori og metode. Roskilde : Roskilde Universitetsforlag,. 216 p.

Ørom, A (2000). Information science, historical changes and social aspects : a Nordic outlook. In: Journal of Documentation, 56 (1). 12 – 26.

Lykke-Nielsen, Domain analysis, an important part of thesaurus construction

Request no.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A. Organisms&testsys | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | 1 | | 2 | | | | | | | | | | | | | | | |
| B. Persons | | | | | | | | | | 2 | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | |
| C. Chemicals & drugs | 1 | 2 | 6 | 1 | 1 | 3 | 1 | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | 1 | 1 | 1 | 1 | | | 1 | 1 | 2 | 2 | 1 | | | 1 | 1 | | 1 | 1 | | 6 | 1 | | 1 | 1 | | | | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| D. Pharmacology | | | | 1 | 1 | | | | | | | | | | | | | | 1 | | 1 | | | 1 | 1 | | | | | | | 1 | 1 | 1 | | | 1 | | 1 | | | | | | | | | | | |
| E1. CNS diseases | 1 | 1 | 1 | 1 | | | | | | 1 | | | | 1 | | | | | | | | | | | | 2 | | 1 | | | | 1 | | | | | | | | | | | | | 1 | | 1 | 1 | | |
| E2. Other diseases | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | 1 | | | | | | | | | | | | | |
| E3. AE & reactions | | | | 1 | | | | | 1 | 1 | | 1 | 1 | 1 | | | | 1 | | | | | 1 | | | | | | | | | | | | | | | | | 1 | | | | | | 1 | | | 1 | |
| F. Therapeutics | 1 | 1 | | | | | | | | | | 1 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| G. Anatomy | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| H. Physiology | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | |
| I. Psychology | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| J. Analytical methods | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 2 | | | | | 1 | | | | | 1 | 1 | | | | | | |
| K. Equipment | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| L. Physics | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| M. Manufacture | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| N. Hygiene | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| O. Working condition | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| P. Management | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | 1 | | | | | | | | | | |
| Q. Quality | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| R. Geographic location | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | | | | |
| S. Languages | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| T. Document types | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| W. Document types | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | | | | | | | | | | | | | | | |
| U. External relations | 1 | | | | | | | | | | | | | | | 1 | | | | | 1 | | | | | 1 | 1 | 1 | 1 | 1 | | | | | | | | | | | | 1 | 1 | | | | | | | |
| V. Units | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

**Appendix A: Requests divided in concept categories**

| Concept type | Frequency in requests (%) |
|---|---|
| C. Chemicals & drugs | 80% |
| E1. Diseases | 32% |
| E3G. Adv. events & reactions | 24% |
| U. External relations | 20% |
| D. Pharmacology | 18% |
| J. Analytical methods | 8% |
| B. Persons | 8% |
| A. Organism & test systems | 4% |
| Q. Quality | 4% |
| E2. Other diseases | 2% |
| T. Document types | 2% |

**Appendix B: Frequency of concept types**

---

*Use/ Used for (USE/UF):* Alternative terms expressing the same concept (synonyms, acronyms).
*Use full name/Acronym (GO/ACR):*
*Use Lu number/Generic name (SEE/GN):*
*Broader term (BT):* Hierarchically superior terms, clarifying the context and meaning of the
      concept and suggesting alternative (more general) terms.
*Narrower term (NT):* Hierarchically inferior terms, clarifying the meaning of the concept and
      suggesting alternative (more specific) terms.
*Related term (RT):* Terms associated or related to the concept, clarifying the context and usage of
      the concept within the company.
*GMP quality system (GMP):* If the term belongs to the GMP quality system terminology, this
      field is activated by adding the source
*GLP quality system (GLP):* If the term belongs to the GLP quality system terminology, this field
      is activated by adding the source
*GCP quality system (GCP):* If the term belongs to the GCP quality system terminology, this field
      is activated by adding the source
Scope note (SN): *Definition of the term. May contain additional information such as drug effects*
      *or disease etiolog.y*
*Subject category (SC):*Each term is classified using a predefined set of subject categories

**Appendix C: General structure of a thesaurus Record**

Lykke-Nielsen, Domain analysis, an important part of thesaurus construction

# Appendix D: Sample of thesaurus records

**Adverse events**
UF: ADE
Adverse drug events
Adverse experiences
Adverse symptoms
Undesirable experiences
Unfavorable signs
Unfavorable symptoms
Unintended signs
Unintended symptoms
Unwanted effects
ACR: AE
NT: Abnormal laboratory findings
Adverse drug reactions
Serious adverse events
Treatment emergent signs or symptoms
Unexpected adverse events
RT: C-T prolongation
Pharmacovigilance
Post-authorization safety studies
Treatment withdrawal
SN: Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product whether or not related to the medicinal (investigational) product. (ICH)
SC: E3 Disease signs, adverse events and reactions

**Animals**
UF: Experimental animals
Laboratory animals
BT: Biological test systems
NT: Dogs
Fishes
Gene manipulated animals
Guinea pigs
Insects
Mice
Rabbits
Rodents
RT: Species
SC: A Organisms and test systems

**Drug substances**
UF: Active components
Active ingredients
Active pharmaceutical ingredients
Active substances
Actives
Agents
Bulk pharmaceutical chemicals
ACR: API
BT: Components
NT: New drug substances
Intermediates
RT: Investigational medicinal products
SN: Any substance or mixture of substances intended to be used in the manufacture of a drug (medicinal) product and that, when used in the production of a drug, becomes an active ingredient of the drug product. Such substances are intended to furnish pharmacological activity or other direct effect in the diagnosis, cure, mitigation, treatment, or prevention of disease or to affect the structure and function of the body. (ICH)
SC: C0 Drugs and chemicals, general

**Lu 00-010**
GN: Cis(Z)-flupentixol
BT: Lu 00-012
NT: Lu 05-110
Lu 08-020
Lu 09-024
Lu 10-040
Lu 10-041
Lu 13-132
SC: C1 Lu drugs

**Lu 10-171**
UF: Celexa (TM)
Cipram (TM)
Cipramil (TM)
Elopram (TM)
Nitalapram
Prisdal (TM)
Seropram (TM)
GN: Citalopram
BT: Serotonin uptake inhibitors
RT: Depression
Forest Laboratories Inc.
Mitsui Pharmaceuticals
Panic disorder
SN: A selective neuronal serotonin reuptake inhibitor and a clinically effective antidepressant w. tolerable side effects. The drug is also effective in reducing ethanol uptake in alcoholics and is used in depressed patients who also suffer fr. tardive dyskinesia (TD) in preference to tricyclic antidepressants, which aggravate this condition. (MESH) CAS Number: 59729-33-8
SC: C1 Lu drugs

**Parkinson's disease**
UF: Parkinson disease
Parkinsonism
Parkinsons
Shaking palsy
RT: Lewy bodies
Lu 00-722
Lu 02-648
Tremor
SN: A progressive, degenerative neurologic disease characterized by a tremor that is maximal at rest, retropulsion (i.e., a tendency to fall backwards), rigidity, stooped posture, slowness of voluntary movements, and a mask-like facial expression. Pathologic features include loss of melanin containing neurons in the substantia nigra and other pigmented nuclei of the brainstem. Lewy bodies are present in the substantia nigra and locus coeruleus. (MESH red)
SC: E1 CNS diseases and disorders

**Subject identification codes**
UF: Patient identification codes
BT: Codes
RT: Confidentiality
Subject identification code lists
Trial subjects
SN: A unique identifier assigned by the principal investigator to each subject in a clinical trial to protect the subjects identity and used in lieu of the subjects name in the investigator reports adverse events and or other trial related data. (ICH)
SC: V Units

**Test systems**
UF: Test subjects
NT: Biological test systems
Chemical test systems
Physical test systems
Plant test systems
SN: Any biological, chemical or physical system or a combination hereof used in a non-clinical study. (OECD)
SC: A Organisms and test systems