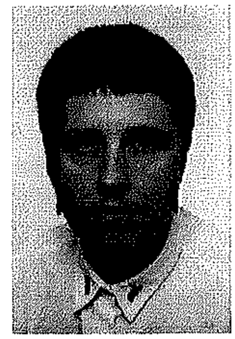
Common Entities and Missing Properties: Similarities and Differences in the Indexing of Concepts

Mirja Iivonen[*] & Katja Kivimäki
Department of Information Studies, University of Oulu, Oulu, Finland

Mirja Iivonen (* To whom all correspondence should be addressed) is a Professor of Information Studies at the University of Oulu, Finland. While writing this article she was working as a visiting scholar at the College of Library and Information Science at the University of Maryland.

Katja Kivimäki has a Masters in Information Studies and currently works as a librarian in the Professional Education Institute in Lapua. She is also working on her Ph.D. and is a student of Professor Iivonen.



Iivonen, M. & Kivimäki, K. (1998). **Common Entities and Missing Properties: Similarities and Differences in the Indexing of Concepts.** *Knowledge Organization, 25(3),* 90-102, 30 refs

ABSTRACT: The selection and representation of concepts in indexing of the same documents in two databases of library and information studies are considered. The authors compare the indexing of 49 documents in KINF and LISA. They focus on the types of concepts presented in indexing, the degree of concept consistency in indexing, and similarities and differences in the indexing of concepts. The largest group of indexed concepts in both databases was the category of entities while concepts belonging to the category of properties were almost missing in both databases. The second largest group of indexed concepts in KINF was the category of activities and in LISA the category of dimensions. Although the concept consistency between KINF and LISA remained rather low and was only 34 %, there were approximately 2.2 concepts per document which were indexed from the same documents in both databases. These common concepts belonged mostly to the category of entities.

Introduction

The purpose of this article is to consider the selection and representation of concepts in indexing. Indexing is one of the basic methods used in knowledge organization. It involves the analysis and description of the contents of documents in such a way that it becomes possible to find them and the messages they contain when searching information. (Anderson, 1995) It is important to emphasize that in indexing two different levels can be found: the conceptual level (analysis) and the terminological level (description).

In book-oriented, or entity-oriented indexing (see Soergel, 1985) at the conceptual level the content of a document is analyzed, and messages and concepts the document contains are recognized. When indexers analyze a document, they do not scan words as strings of characters but as signs of concepts, ideas which the document discusses. Indexing starts with an intellectual analysis of a document. Similarly, as the intellectual work in general, the intellectual analysis of documents also includes some degree of interpretation, involving different understanding of signs.

After analyzing documents, indexers must move from the conceptual level to the terminological level. They must express the concepts they recognized with words, e.g. with index terms. The expression of concepts always requires a term, and it is impossible to talk about concepts without terms.

In request-oriented indexing (see Soergel, 1985) the content of a document is compared to the terms of a controlled vocabulary and indexers are deciding with which words the document should be found. Also in request-oriented indexing both the conceptual and terminological level can be recognized. Indexers do not compare the words just as strings of characters but as signs of concepts. At the conceptual level they consider topics which should be found in searching documents. At the terminological level they express the results of the comparison with words.

Although there are many rules and procedures for indexing, there is also much variety in this process. There are many studies which show how inconsistently indexers use words in describing the same documents (see e.g. Lancaster, 1968, Zunde & Dexter, 1968, Funk, Reid & McCoogan, 1983, Markey, 1984, Iivonen 1989, Iivonen, 1990, Lancaster, 1991.) Other studies show that also searchers use various words in searching the same topic (see e.g. Fidel, 1985, Fidel, 1987, Saracevic & Kantor, 1988, Iivonen, 1995a, Iivonen, 1995b). On the other hand, previous studies (Saracevic, 1984, Iivonen 1989, Iivonen, 1990, Iivonen, 1995a, Iivonen, 1995b) show also that both indexers and searchers are more consistent in the selection of concepts than in the selection of index/search terms. Although indexers and searchers often use various words in describing the same documents or search requests, they still refer to the same concepts. In other words, they talk about the same thing with different words. An interesting and still open question is what are those things or concepts indexers and searchers are talking about. We suppose that it is easier for them to recognize and describe certain kinds of concepts and leave others out. Unfortunately most studies of indexing focus on terms, not concepts, and so do also most rules for indexing.

In this article we focus on concepts in indexing – the topic which has often stayed in the shadow of words. We pay attention to the types of concepts selected in indexing to tell about the content of documents. In addition we consider the consistency, i.e. the degree of agreement, in the selection of concepts in indexing of the same documents on various occasions.

Concepts

A concept is a knowledge unit, which combines a referent and a linguist symbol or a verbal form (a sign) that is used in describing it. The linguist symbol is the name of the concept, and the concept is the meaning of the linguist symbol, the internal image with which the phenomena of the external world can be classified. (Dahlberg 1978, p.143-144, Dahlberg 1980, p.217, Fugmann 1982, p.141, Fugmann 1985, p.117, Fugmann 1993, p.17-18.)

The relationship between linguist symbol, concept and referent is usually described with the semantic triangle as shown in Figure 1.

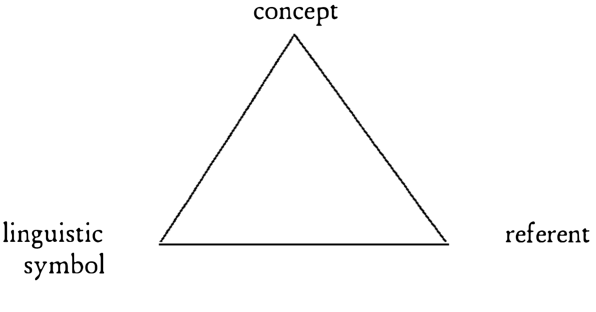


Figure 1: Semantic triangle

Various concepts can be arranged into categories. Aristotle (quoted here from the 1994 Finnish edition, p.7-30) postulated the following basic categories: *substance, quantity, quality, relations, time, space, position, possession, activities* and *objects*. Dahlberg (1978, p.145 and 1981, p.20) has used this classification as the basis for her own system of categories. According to Dahlberg concepts can represent *entities*, i.e. both material objects, immaterial objects and principles; *properties*, i.e. quantities, qualities and relations; *activities*, i.e. processes, operations and states; and *dimensions*, i.e. time, space and positions.

According to Ranganathan (1957, p.168-185) basic categories are *personality, matter, energy, space* and *time*. The categories of matter, space and time are clear. The category of energy includes activities, processes and problems. The category of personality is more related to the context and may include both things, kinds of things or actions and kind of actions, depending on the bias of the subject field. Aitchison and Gilchrist (1987) have used Ranganathan's system as a basis for their own system. According to them, concepts can be divided into the following categories: entities, actions, time and space. In their system, the category of entities includes also properties. That is different from Dahlberg's system, where the category of properties is presented as its own basic category. Another difference is that Dahlberg uses the category of dimensions where she, in addition to time and space, also has concepts representing positions. In other respects Dahlberg's, Aitchicon's, and Gilchrist's systems are very close to each other.

The basic categories of concepts differ clearly from each other. It is possible that some types of concepts might be more easily and consistently recognized than others. According to Soergel (1985, 225-249) it is possible that in request-oriented indexing more attention is paid to abstract concepts than in entity-oriented indexing.

According to Dahlberg (1978, 147) concepts can be divided into three levels according to their specificity. These levels *are general concepts, special concepts* and *individual concepts*. A referent of a general concept refers to all items of a given kind. A referent of a special concept refers to some items of a given kind. A referent of an individual concept refers only to one item of a given kind. For example, the linguist symbol *libraries* stands for a general concept and refers to all libraries. The linguist symbol *public libraries* stands for a special concept and refers to only some kind of libraries. The linguist symbol *City library of Oulu* stands for an individual concept and refers to one library only. Iyer (1995, p.47) states that in a hierarchy there is always a certain basic level of concepts, where ideas will concentrate and where concepts are more familiar to people than concepts at other levels. Iyer's ideas are based on Rosch's (quoted here from Lakoff, 1987) prototype theory according to which there are basic-level categories and knowledge is mainly organized at this basic level. The basic level concepts are easier to recognize and visualize[1] than other forms. Because indexing is a knowledge organization process, it is possible that also in indexing a certain hierarchical level of concepts is selected more often than others.

Research Questions

In this article we consider and compare the indexing of concepts in two databases of library and information studies, KINF and LISA. We address the following research questions:

1. Which types of concepts are most often presented in indexing in KINF and LISA?

2. What is the degree of concept consistency when the same document is indexed in KINF and LISA? The concept consistency means the degree of agreement in the selection of concepts on various occasions.

3. Which kind of differences occurs in the indexing of concepts in KINF and LISA?

Data and Methods

The data for this study was collected in spring 1997 in KINF and LISA. KINF is a Finnish database of library and information studies. LISA is an international database of library and information studies. The indexing in KINF is carried out with a Finnish vocabulary and in LISA with an English vocabulary. Because the purpose of the study was to compare indexing of the same documents in two databases, one of the authors (Kivimäki) sought in both databases bibliographic citations of all documents, which were published in Finland either in Finnish in 1995 or in English in 1995-1996. 38 Finnish and 11 English documents were found for the research data. All these documents were articles. In the research data there were altogether bibliographical citations to 49 documents with total 287 index terms or descriptors in KINF and 207 index terms or descriptors in LISA (see Table 1).

To be able to recognize concepts represented in indexing, index terms (=linguist symbols) were considered. Because KINF uses precordinated indexing while LISA does not, precordinated index terms in KINF were first broken down. For example a precordinated index term *kirjastonhoitajat : koulutus* (librarians : education, colon was used for precordination) was broken down into the components *librarians* and *education.*

To recognize concepts used in indexing, Dahlberg's system of categories of concepts was used as a basis of analysis. The following basic categories of concepts were used:

1) entities, i.e. objects and things,

2) properties, i.e. quantities, qualities and relations,

3) activities, i.e. operations, processes and states, and

4) dimensions, i.e. concepts related to time and space.

As entities we took into account:

i) concrete objects (Dahlberg's material objects), which were living beings that were calculable, e.g. *librarians, customers, researchers;*

ii) concrete objects (Dahlberg's material objects), which were lifeless material things which were calculable, e.g. *databases, catalogues, cd-roms;*

iii) concrete objects (Dahlberg's material objects), which were not calculable, e.g. *information technology, staff;*

iv) abstract things (Dahlberg's immaterial objects) and their systems, e.g. *knowledge, learning culture, collection policy;*

v) abstract-concrete systems representing communities and organizations, e.g. *European Union, universities, libraries.*

As properties only qualities occurred in the data, when the concept *modernity* was found. Quantities and relations were not found in the data.

As activities we took into account:

i) operations, which were rather simple and routine activities, or transitive operations which had the object of the action, e.g. *reporting, lending;*

ii) processes, which were intra-actor and often intransitive activities, e.g. *development, learning, information seeking;*

Table 1. Research data

	Bibliographical citations			Index terms/Descriptors		
Database	Finnish documents	English documents	Σ	Finnish documents	English documents	Σ
KINF	38	11	49	238	49	287
LISA	38	11	49	160	47	207

iii) states, which were complex and composing activities and might include both operations and processes and which do not have only one subject, e.g. *knowledge management, knowledge work, electronic publishing.*

As dimensions only space occurred in the data. They were concepts representing geographical locations, e.g. *Oulu, Finland.* Concepts of time were not found in the data.

When the concept-consistency ($CC_{1\leftrightarrow 2}$) of KINF and LISA was calculated, the following formula[2] (Lancaster, 1968) was used

$$CC_{1\leftrightarrow 2} = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

In the formula C_1 means the set of concepts presented in KINF and C_2 means the set of concepts presented in LISA.

For calculating concept consistency, different index terms were counted as the same concept in the following situations:

1. if there was a difference only in the language used (one term was in English and another one in Finnish)
   KINF: *Kirjastot* (in English *Libraries)*
   LISA: *Libraries*
2. if index terms were the singular and plural forms of the same term
   KINF: *Eurotietokeskus* (singular form)
   LISA: *European documentation centres* (plural form)
3. if there was an equivalence relationship between the terms
   KINF: *Pohjoismaat* (in English *Nordic countries); Kokoukset* (in English *Meetings)*
   LISA: *Scandinavia; Conferences*
4. if there was a hierarchical relationship[3] between the terms. In these cases the concepts between which there was a hierarchical relationship belonged into the same concept category (e.g. the category of entities)
   KINF: *Tietokannat* (in English *Databases); Erikoiskirjastot* (in English *Special libraries)*
   LISA: *Online databases; Business libraries*

If the same concept was indexed in the same database with two or more index terms between which there was an equivalence *(e.g. Nordic countries - Scandinavia)* or a hierarchical relationship *(e.g. libraries - special libraries),* they were counted as one concept both in calculating consistency and in calculating the numbers of concepts used in indexing.

Results

Number of Concepts and Concept Consistency

In indexing only a few concepts of documents will be represented. Completely exhaustive indexing is practically impossible - indexers cannot repeat the whole content of a document but have to make choices. They have to decide which concepts should be indexed so that the document could be found in searching. These decisions are related to the level of indexing exhaustivity and the indexing policy of databases (or other document collections).

In our study only four or five concepts per document were represented in indexing, in KINF on average 4.8 concepts per document and in LISA on average 3.6 concepts per document (see Table 2). The number of concepts is very near to the number of concepts indexed in Iivonen's (1989) study. In her study 10 indexers indexed the same ten documents (monographs) and on average 3.5 concepts per document were presented. In indexing the number of concepts to be presented seems to remain rather low. We can assume that because indexers select only a few concepts per document for indexing they usually aim to select core concepts to tell about the main ideas discussed in the document.

Table 2. Number of concepts represented in indexing

Database	*Finnish documents* Concepts/document Mean value	*English documents* Concepts/document Mean value	*All documents* Concepts/document Mean value
KINF	5.1	4.0	4.8
LISA	3.7	3.3	3.6
KINF+LISA	6.7	5.6	6.3

Some of the indexed concepts in our data were common concepts and represented in both databases (see Table 3), while some were presented only in one database. When we took into account both databases simultaneously, approximately six concepts per document were indexed altogether (see Table 2). Indexing the same document on various occasions seems to add the number of concepts to be selected for indexing from the same document and will give more access points to the content of this document.

There were some differences in the number of concepts when the same documents were indexed for two databases (see Table 2). More concepts were presented in KINF than in LISA. We can find at least two reasons for this. First, our data consisted of documents published in Finland either in Finnish or in English. The role of Finnish documents is more crucial in KINF (Finnish database) than in LISA (international database) and probably they were therefore indexed for KINF more exhaustively than for LISA. Second, we can assume that because of the language it was easier for the indexers of KINF than for the indexers of LISA to interpret the content of documents and recognize their main concepts. The difference in the number of concepts between KINF and LISA was larger in indexing documents published in Finnish than in indexing documents published in English.[4] The number of concepts presented in indexing is not constant but depends at least on the following factors. First, it depends on the importance of a document to the environment (e.g. database, library collection) for which it will be indexed. Second, it depends on the ease of understanding (e.g. language) and interpreting the document.

The concept consistency remained rather low. The mean value was only 34 % (see Table 3). There was, however, some consistency in indexing of concepts in KINF and LISA. Some concepts were indexed in both databases. This finding supports the idea that the main content (concepts) of documents will be indexed consistently (Iivonen 1989, Iivonen 1990, Lancaster 1991). On the other hand, some concepts were indexed only in KINF or in LISA. Two examples of differences in indexing of the same documents in KINF and LISA are presented in Table 4. One indexed document considered the role of public libraries as information source and was published in English. More concepts were selected for KINF than for LISA. Only one concept *(public libraries)* was chosen for both databases. We can consider *public libraries* as a core concept of this first document. Another document considered the role of the Internet in researchers' work. Four concepts were selected for indexing both in KINF and LISA but only one of them *(electronic publishing)* was a common concept. We can consider *electronic publishing* as a core concept of this second document.

Categories of Indexed and Missing Concepts

We assumed that certain types of concepts might be more easily and more often recognized and presented in indexing than others. For human beings abstract things are usually more difficult to see and understand than concrete things. Soergel's (1985) distinction between entity-oriented and request-oriented indexing (see above) indicates that this happens also in indexing. According to Soergel entity-oriented indexers focus more on concrete than abstract concepts in indexing.

We found differences between the categories of concepts presented in indexing (see Table 5). The clearly largest group of concepts was entities (in KINF 65 % and in LISA 60 %). The second largest group of concepts was in KINF activities (27 %) but in LISA dimensions (21 %). A concept belonging to the properties category was found only in KINF and only once. In indexing the focus seems to be on objects and things about which the documents talk. Activities are difficult to imagine without a subject or an object, and this might be the reason why the focus in indexing is on entities, not on activities. Dimensions form a frame for various entities and activities and can be either temporal or spatial. Dimensions might be

Table 3. Common concepts and concept consistency in KINF and LISA

	Finnish documents	*English documents*	*All documents*
Common concepts/ documents	2.3	1.6	2.2
Concept consistency	36 %	29 %	34 %

Table 4. Examples of differences in indexing in KINF and LISA

Document 1: Tuominen, Kimmo: The public library as information source: Findings of an interview study. - Scandinavian Public Library Quarterly 29 (1): 8-10, 1996.

KINF		LISA	
Index terms	Indexed concepts	Index terms	Indexed concepts
Yleiset kirjastot (=Public libraries)	A	Libraries	A*
		Public libraries	A*
Tiedontarve (=Information need)	B		
Tiedonhankinta (=Information seeking)	C		
Arkielämä (=Everyday life)	D		
Kirjastonkäyttö (=Library use)	E		
		Finland	F**
		Salo	F**
		User surveys	G

Document 2: Ylikoski, Petri: Internet tutkijan apuna (Internet as a researcher's assistance). Signum 28 (7):148-150, 1995.

KINF		LISA	
Index terms	Indexed concepts	Index terms	Indexed concepts
Internet	A***		
Tutkijat (Researchers)	B		
Kirjastot (Libraries)	C		
Elektroninen julkaisutoiminta (Electronic publishing)	D***	Electronic publishing	
	D***	Electronic mail	E***
		Research methods	F
		Information communication	G

* There is a hierarchical relationship between Libraries and Public libraries and therefore these two index terms were counted as one concept.

** There is a hierarchical relationship between Finland and Salo (Salo is a city in Finland) and therefore these two index terms were counted as one concept.

***We interpreted that there are partitive relationships but not analytic partitive relationships between electronic mail and Internet and between electronic publishing and Internet. Therefore there were understood as associative not hierarchical relationships and concepts were regarded as separate concepts.

easy to recognize also for indexing although this is not always done. On the other hand, some documents do not discuss temporal and spatial aspects at all, and in those cases neither dimension can be presented in indexing. The category of properties seems to be a black hole in indexing. Such concepts are almost always ignored.

It was interesting to notice that in the category of entities concrete objects were not indexed more often than abstract things or abstract-concrete systems. However, we were able to see again differences between indexing in KINF and LISA (see Table 5). In KINF the proportion of abstract things (e.g. *information need, everyday life*) was slightly higher than the proportion of concrete objects (e.g. *librarians, databases*). In LISA the proportion of abstract-concrete systems (the concepts represented organizations and communities, e.g. *libraries, European Union*) was higher than the proportion of concrete objects. Instead the proportion of abstract things was only 10 %

Table 5. Categories of indexed concepts in KINF and LISA

	Concepts indexed in KINF						Concepts indexed in LISA					
Categories of concepts	Finnish documents n=193 %		English documents n=44 %		Σ n=237 %		Finnish documents n=142 %		English documents n=36 %		Σ n=178 %	
Entities	*64*		*66*		*65*		*61*		*56*		*60*	
Abstract things		21		27		22		10		11		10
Concrete objects		22		23		22		20		22		20
Abstract-concrete systems		21		16		20		32		22		30
Properties	*1*		-		-		-		-		-	
Quantity		-		-		-		-		-		-
Quality		1		-		-		-		-		-
Relations		-		-		-		-		-		-
Activities	*27*		*23*		*27*		*19*		*19*		*19*	
Operations		7		9		7		3		6		4
Processes		8		9		8		8		3		7
States		12		5		11		8		11		8
Dimensions	*8*		*11*		*8*		*20*		*25*		*21*	
Time		-		-		-		-		-		-
Space		8		11		8		20		25		21
Σ	*100*	100	*100*	100	*100*	98	*100*	101	*100*	100	*100*	100

in LISA and clearly less than in KINF. We can conclude that indexers do not pay attention only to concrete concepts but also take into account abstract concepts, although indexing for LISA seems to be more concrete-oriented than indexing for KINF.

Concepts belonging to the activities category were indexed more often for KINF than LISA. Inside this category concepts representing different states (e.g. *knowledge management*, *library use)* were indexed in both databases more often than concepts representing operations (e.g. *lending*) or processes (e.g. *information seeking*). The finding that rather large proportion of indexed concepts represented activities (in KINF 26 % of concepts and in LISA 19 % of concepts) confirms the idea that indexers do not only look for concrete objects from documents but try to recognize and describe documents' contents from different viewpoints. These different viewpoints add to the value of subject control and might be lost in automatic indexing where only the frequencies of words are calculated. Some viewpoints, e.g. some activities discussed in a document, might be worth indexing for users even if as words (strings of characters) they occur only once in a document.

There was a big difference in indexing concepts representing dimensions between KINF and LISA. In both databases only concepts representing space were indexed, but in LISA more often than in KINF (see Table 5). The explanation for this may be rather simple. LISA is an international database and includes bibliographic citations of documents published in many countries and in many languages. In LISA it is essential to tell users what is the local area with which a document is dealing, if this information can be found in the document. In our data most of the space-concepts in LISA represented the concept *Finland*. (Our data included documents that were published in Finland). KINF is a Finnish database and maybe in indexing documents treating Finland for KINF the concept *Finland* is not considered to be so important. The concept *Finland* does not sort out data in KINF as well as it does in LISA. The role of concepts belonging to the dimensions category seems to be different in various environments (e.g. various databases).

The common concepts belonged mostly (in 68 % of cases) to the category of entities (see Table 6). In most cases they represented either concrete objects or abstract-concrete systems. Instead, the proportion of abstract things (also entities) was not so large in the group of common concepts (only 11 %). Similarly there were not so many concepts among common concepts belonging to the categories of activities or dimensions and no concept belonging to the category of properties. The indexing of concrete objects and abstract-concrete systems might increase the indexing consistency.

There was one interesting difference in common concepts selected from Finnish and English documents. The proportion of dimensions among common concepts was larger when documents were published in English than when documents were published in Finnish. In most cases this common concept was *Finland*. The amount of common concepts selected from English documents was, however, rather small: 18 common concepts in indexing 11 documents.

To be able to find possible shortcomings of concept indexing we paid special attention to missing concepts, ones that were indexed in one database but not in the other. We can assume that these concepts appeared in documents and were at least somehow important concepts because they were indexed in one of the databases[5].

Table 6. Categories of common concepts in KINF and LISA

	Common concepts					
Categories of concepts	Finnish documents n=88 %		English documents n=18 %		Σ n=106 %	
Entities	*70*		*55*		*68*	
Abstract things		11		11		11
Concrete objects		23		22		23
Abstract-concrete systems		36		22		34
Properties	-		-		-	
Quantity		-		-		-
Quality		-		-		-
Relations		-		-		-
Activities	*21*		*17*		*20*	
Operations		5		6		5
Processes		5		6		5
States		11		6		10
Dimensions	*9*		*28*		*12*	
Time		-		-		-
Space		9		28		12
Σ	*100*	100	*100*	101	*100*	100

The categories of missing concepts varied in KINF and LISA (see Table 7). In KINF the largest group of missing concepts were dimensions, especially concepts representing space. We discussed above possible reasons for that. Although these reasons are easy to understand, they are not necessarily reasonable. The ignored dimensions in indexing decrease always the exhaustivity of indexing. It has also an impact on search results. In information retrieval, dimensions are useful concepts in limiting the search.

Table 7. Categories of missing concepts in KINF and LISA

	Concepts missing in KINF but indexed in LISA						Concepts missing in LISA but indexed in KINF				missing	
Categories of concepts	Finnish documents		English documents		Σ		Finnish documents		English documents		Σ	
	n=32 %		n=14 %		n=46 %		n=62 %		n=17 %		n=79 %	
Entities	*19*		*57*		*31*		*68*		*71*		*68*	
Abstract things		9		14		11		32		47		35
Concrete objects		3		21		9		19		18		20
Abstract-concrete systems		6		21		11		16		6		13
Properties	-		-		-		*2*		-		*1*	
Quantity		-		-		-		-		-		-
Quality		-		-		-		2		-		1
Relations		-		-		-		-		-		-
Activities	*22*		*7*		*17*		*26*		*30*		*27*	
Operations		3		7		4		10		12		10
Processes		6		-		4		8		12		9
States		13		-		9		8		6		8
Dimensions	*59*		*36*		*52*		*5*		-		*4*	
Time		-		-		-		-		-		-
Space		59		36		52		5		-		4
Σ	*101*	100	*100*	99	*100*	100	*101*	100	*100*	100	*100*	100

In LISA the largest group of missing concepts was entities and especially concepts representing abstract things. The second largest group of missing concepts was activities. Both abstract things (e.g. *information need, everyday life*) and activities (e.g. *information seeking, learning*) are intangible, unlike concrete objects (e.g. *books, librarians*), and they require abstract thinking and ability to identify various phenomena. As Soergel (1985) has pointed out, entity-oriented indexers seem to pay attention to concrete concepts while search request-oriented indexers are looking for abstract concepts. Maybe the indexing policy of LISA focuses on the indexing of concrete objects. This means, however, that only the surface of the document will be indexed and many aspects that are useful to searchers will be ignored.

The Indexing of Individual Concepts

A linguist symbol of an individual concept refers only to a *single* item of a given kind while a linguist symbol of a special concept refers to *some* items of a given kind and a linguist symbol of a general concept refers to *all* items of a given kind. The linguist symbol of an individual concept is a proper name (e.g. *European Union*, *Salo* or *Finland)*, but special concepts and general concepts are expressed by common nouns (e.g. *organizations*, *cities* or *countries*). The use of proper names as index terms will sort documents rather well but is not done very consistently in indexing.

We found clear differences between KINF and LISA in the use of individual concepts in indexing (see Table 8). The proportion of individual concepts from all indexed concepts was much higher (41 %) in LISA than in KINF (24 %). We noticed already earlier that in LISA dimensions were indexed in many cases and that in most cases those dimensions were concepts representing space, especially the concept *Finland*. When we excluded geographical concepts from the group of individual concepts presented in KINF and LISA, the differences between databases were not so large any more. Now the amount of individual concepts (other than geographical concepts) in KINF was 16 % from all concepts and in LISA 18 % from all concepts.

Individual concepts can be indexed either alone or with general or special concepts. When an individual concept is indexed with general or special concepts, indexers will give additional information and tell users into which group the individual concept belongs. For example, when some special library, e.g. *Eduskunnan kirjasto* (The Library of Finnish Parliament) is indexed, indexers may want to express also the special concept *special libraries* and/or the general concept *libraries*. Instead, when an individual concept is indexed alone, indexers do not see it necessary to state into which group the individual concept belongs, e.g. indexers will present an individual concept *Salo* but not a general concept *cities*, or an individual concept *Finland* but not a general concept *countries*. A rule given many times to indexers is to select as specific level of index terms/concepts as possible and use it. This does not mean, however, that they should use only this specific level.

Table 8. The use of individual concepts in KINF and LISA

	Concepts indexed in KINF			Concepts indexed in LISA		
	Finnish documents	English documents	Σ	Finnish documents	English documents	Σ
	n=193	n=44	n=237	n=142	n=36	n=178
	%	%	%	%	%	%
Individual concept was used						
*alone	17	11	16	35	31	34
* with general or special concept	9	2	8	6	8	7
Individual concept, excluding geographical concepts, was used						
*alone	9	-	8	13	3	11
*with general or special concept	9	2	8	6	8	7

In our data, individual concepts were used alone when they were geographical concepts *(Finland* and *Ireland)* but in other cases either alone (e.g. *Europe Institute* and *European Union*) or with more general concept (e.g. *LINNEA* (proper name) with *national networks* and *Nordic Centre of Excellence for Electronic Publishing* with *information centres*) (see Table 8). We noticed also that in KINF individual concepts were presented with special/general concepts slightly more often than in LISA. Maybe one reason for this is that the indexing of KINF was more exhaustive than the indexing of LISA and this affected also the selection of general concepts together with individual concepts.

Discussion

In this study the indexing of the same documents in two various databases was compared. One of the databases (KINF) is a national database serving rather limited clientele and using for indexing the vocabulary in a rather rare language (Finnish). The other database (LISA) is an international database serving worldwide clientele and using for indexing its own vocabulary and almost a universal language (English). The differences of indexing we found might be partly the result of using different vocabularies and partly of the different indexing practices. Indexing practice can be described as those work routines, procedures, rules, and restrictions followed in the indexing process. It includes both indexing policy and specific decisions made by indexers (Iivonen and Sonnenwald, 1998). From the user's point of view the key question is which topics and concepts are represented in indexing and which are not. For them, the question of secondary importance is why certain topics and concepts can or cannot be found in searching. They are not concerned whether this happens because of the vocabulary or indexing practice.

Although there were clear differences in indexing of concepts in KINF and LISA, also some common concepts selected for both databases were found. In KINF on average 4.8 and in LISA on average 3.6 concepts per document were indexed. 2.2 of them were common concepts represented in both databases. We can assume that the main content of a document will be indexed rather consistently although the indexing will take place in various environments for various clienteles. We can regard the indexing as a special kind of expertise where indexers know how to compress the content of a document into a few main concepts. This compression requires the understanding and interpretation of documents and cannot be done merely by calculating the frequencies of words.

Although the clear concept-consistency was found in the indexing for KINF and LISA, some inconsistency was also found. The mean value of concept-consistency was only 34 %. The differences in indexing were in two directions, on one hand in the exhaustivity of indexing and on the other hand in the categories of indexed concepts.

The indexing in KINF was more exhaustive than the indexing in LISA. That means that more concepts of documents were presented in KINF than in LISA. This is understandable because all documents in our data were published in Finland, either in Finnish or in English. As a national database KINF offered many points of views to Finnish documents while LISA described them more cursorily. We can assume that those documents are more important for indexers in the "near neighborhood" than at the international level, especially if we consider documents of the small area (either geographical or linguistic area).

Most concepts that were indexed both in KINF and LISA belonged to the category of entities. Activities and dimensions were rarely selected, and properties were almost totally missing in the indexing. The entities were also the clearly largest group of common concepts presented in both databases. The indexing seems to focus on entities that are also easier to identify consistently from documents than on other types of concepts. Because users are, however, interested also in other aspects of documents, indexers should consciously try to seek and describe activities, dimensions and properties discussed in documents.

There were differences in the indexed concepts between KINF and LISA. More abstract things (a subcategory of entities) and activities were presented in KINF than in LISA while more dimensions were presented in LISA. The indexing in KINF seems to be more abstract-oriented than the indexing in LISA. The identification and presentation of abstract things are not always very easy but serve users who are interested also in abstract issues (e.g. *information need, knowledge, learning culture*) and not only in concrete things (e.g. *books, librarians*).

More individual concepts were presented in LISA than in KINF. In most cases they were geographical concepts belonging to the category of dimensions, and in many cases the concept was *Finland*. This concept definitely sorts out documents conveniently in the international environment. Individual concepts in KINF were combined with more general concept more often than in LISA. The indexing of the same concept at various levels of hierarchy offers more access points to the same concept. Different users may approach documents at different levels of hierarchy and still find the same document useful.

Conclusion

Indexing has an essential role in knowledge work because it organizes the content of documents and allows users to find messages that documents carry. Indexing is a value-adding process, and serves users, although some indexers might see indexing as one kind of an art in itself. It is for the benefit of users that principles and practices of indexing are studied. In this article we considered the selection of concepts in indexing, the research area which has remained in the shadow of words used in indexing. Users are more interested in meanings and messages that documents discuss than in words as strings of characters. It is necessary therefore to pay attention also to concepts presented in indexing. We found a certain degree of consistency in the indexing of concepts, but we found also clear differences when the same documents were indexed for various databases. Our findings could be taken into the consideration when vocabularies will be updated and indexing policies will be reformulated.

Notes

1. When people are asked to list things, they mostly list things at the basic level (chair, car, dog), not at the superordinate level (furniture, vehicle, mammal) or subordinate level (rocking chair, sport car, retriever) (Lakoff, 1987).
2. There are many different formulae for calculating consistency. Various formulae give different consistency data points. Therefore the comparison of the results of consistency studies is difficult, many times even impossible. (Iivonen, 1993).
3. For hierarchical relationships we took into account generic relationships, instance relationships between individual concept and its general concept and analytic partitive relationships. Other partitive relationships were excluded as associative relationships. There is an analytic partitive relationship be

tween terms/concepts if we can say: What is said to be true of a part is true of its whole (Hutchins, 1978.)

4. Because we did not receive the answer from the producers of LISA, we do not know but we can guess that the indexing of Finnish documents for LISA might have been done on the basis of English abstracts.
5. Of course in documents there appeared also many other concepts which were missing in indexing for both databases but we ignored them in this article.

Acknowledgements

We thank Dagobert Soergel and Rebecca Green for their useful comments on an earlier version of this article.

References

Aristoteles (1994). Kategoriat. In *Aristoteles II: kategoriat. Tulkinnasta. Ensimmäinen analytiikka.* Helsinki: Gaudeamus. 7-38.

Aitchison, J. & Gilchrist, A. (1987). *Thesaurus construction. A practical manual.* 2. ed. London: Aslib.

Anderson, J. D. (1997). Organization of knowledge. In J. Feather & P. Sturges (Ed.). *International Encyclopedia of Information and Library Science.* London: Routledge. 336-353.

Dahlberg, I. (1978). A referent-oriented, analytical concept theory for INTERCONCEPT. *International Classification 5 (3).* 143-151.

Dahlberg, I. (1980). New trends in classification. In P. J. Taylor (Ed.). *New trends in documentation and information. Proceedings of the 39th FID Congress.* London: Aslib. 214-222.

Dahlberg, I. (1981). Conceptual definitions for INTERCONCEPT. *International Classification 8 (1).* 16-22.

Fidel, R. (1985). Individual variability in online searching behavior. In C. A. Parkhurst (Ed.) *ASIS '85: Proceedings of the American Society for Information Science 48th Annual Meeting.* White Plains, NY: Knowledge Industry Publications. 69-72.

Fidel, R. (1987). What is missing in research about online searching behavior. *Canadian Journal of Information Science 12 (3/4).* 54-61.

Fugmann, R. (1982). The complementary of natural and indexing languages. *International Classification 9 (3).* 140-144.

Fugmann, R. (1985). The five-axiom theory of indexing and information supply. *Journal of the American Society for Information Science 36 (2).* 116-129.

Fugmann, R. (1992). Indexing quality: Predictability versus consistency. *International Classification 19 (1).* 20-21.

Fugmann, R. (1993). *Subject analysis and indexing: Theoretical foundation and practical advice.* Frankfurt/Main: Indeks Verlag.

Fugmann, R. (1994). Representational predictability: Key to the resolution of several pending issues in indexing and information supply. In *Knowledge organisation and quality management.* Ed. Hanne Albrechtsen and Susanne Oernager, p. 414-422. Frankfurt/Main: Indeks Verlag.

Funk, M. E., Reid, C. A. & McCoogan, L. S. (1983). Indexing consistency in medline. *Bulletin of Medical Library Associations, 71 (2).* 176-183.

Hutchins, W. J. (1978). *Languages of indexing and classification.* Stevenage: Peter Peregrinus.

Iivonen, M. (1989). *Indeksointituloksen riippuvuus indeksointiympäristöstä.* Tampereen yliopist•, Kirjastotieteen ja informatiikan laitoksen tutkimuksia 26. Tampere: Tampereen yliopisto.

Iivonen, M. (1990). Interindexer consistency and the indexing environment. *International Forum on Information and Documentation 15 (2).* 16-21.

Iivonen, M. (1993). Johdonmukaisuuden laskeminen tiedon tallennuksen ja haun tutkimuksessa. *Kirjastotiede ja informatiikka 12 (2).* 63-76.

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

Iivonen, M. (1995b). *Hakulausekkeiden muotoilun yhdenmukaisuus onlineviitehaussa.* Tampere: University of Tampere. (With English summary)

Iivonen, M., Sonnenwald D. H. (1998). From translation to navigation of different discourses: A model of search term selection during the pre-online stage of the search process. *Journal of the American Society for Information Science 49 (4).* 312-326.

Iyer, H. (1995). *Classificatory structures: Concepts, relations, and representation.* Frankfurt/Main: Indeks Verlag.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: The University of Chicago Press.

Lancaster, F. W. (1968). *Evaluation of the Medlars demand search service.* Washington D.C.: National Library of Medicine.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice.* London: The Library Association.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research 6 (2).* 155-177.

Ranganathan, S. R. (1957). *Prolegomena to library classification.* 2. ed. London: Library Association.

Saracevic, T. (1984). Measuring the degree of agreement between searchers. In B. Flood, J. Witiak &

T. H. Hogan (Ed.). *ASIS '84: Proceedings of the American Society for Information Science 47th Annual Meeting.* White Plains, NY: Knowledge Industry Publications. 227-230.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science 39 (3).* 197-216.

Soergel, D. (1985). *Organizing information: Principles of database and retrieval systems.* San Diego: Academic Press.

Zunde, P. & Dexter, M. E. (1969). Indexing consistency and quality. *American Documentation 20 (4).* 259-267.

Mirja Iivonen, University of Oulu, Department of Information Studies, POB 111, 90571 Oulu, Finland

Katja Kivimäki, University of Oulu, Department of Information Studies, POB 111, 90571 Oulu, Finland