Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY.
6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE FOR INDEXING IN BIOTECHNOLOGY.

Sara von Ungern-Sternberg, PhD[1]
Åbo Akademi University, Department of Library and Information Science

Abstract

Interdisciplinary research crosses by definition the boundaries between fields of research. This creates difficulties for the indexing of interdisciplinary works since indexing languages follow the establiched divisions and definitions of subject fields. For the information specialist and the librarian it is still necessary to find methods for the indexing of interdisciplinary fields in order to carry out the tasks of information provision and collection development. Presented is a method to identify the emerging indexing language in a new interdisciplinary field. The method is based on the comparison and analysis of the indexing of articles co-occuring in different bibliographic databases.

The study reported here is part of a larger project (von Ungern-Sternberg, 1994) addressing issues for collection development in interdisciplinary fields. The aim of this study is to show that it is possible to gather terms for an indexing language on the basis of the indexing done for bibliographic databases.

One of the difficulties in developing collections in interdisciplinary fields is the weaknesses of the existing documentation languages (the classification systems and indexing languages); they often reflect an obsolete division of the subject fields. Still the development of any subject field depends on how well it can make use of its knowledge base. To identify and organise information sources, for instance the production of literature, are therefore important issues in the development of subject fields. How can the librarian decide which material should be included in the collections when even the scientists of the field have little concensus about the definition of the field? How can the librarian build up collections in fields where the classification systems, used for instance by journal services, have a discipline based structure and including interdisciplinary subject fields only at a very basic level, if at all. What methods are there to define a documentation language which is up to date and reflects the content of local research? These were the research questions for the project. One of the goals was to find a method for selecting terms for a documentation language in such fields. In this study the terminology in an interdisciplinary field was analysed in order to develop terms for a documentation language. This language should build on contemporary research and thus reflect the information needs in the field.

Another goal was to define the stage of research in an interdisciplinary field through the documentation language. Trofimenko has developed a method for this through index terms. According to him the number of low frequent terms describes the renewing of the field and reflects searching for new research methods and problems in the field. At initial stages of research, when the number of terms is small, different terms are used to define a search topic. Therefore a database

[1] formerly Sara von Flittner

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY.
6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

mostly consists of low-frequency terms. When the research level rises, the part of low and high frequent terms goes down, but the part of mean frequent terms rises. Low, and high frequent terms can be associated with research, while the mean frequent terms are associated with the technical implementation of the research results (Trofimenko, 1990).

TERMS FOR A DOCUMENTATION LANGUAGE

How are terms gathered for a documentation language in a specific subject field? According to Beghtol (1986) there are four basic types of warrant of terms: literary warrant, scientific/philosophical warrant, teaching warrant and cultural warrant. The different types together form the unique semantic theory and the characteristics of any classification system. The concept literary warrant was introduced by Hulme in 1911 in his book "Principles of Book Classification". According to him a class in a classification scheme is justified only if there is at least one book in the field. In other words a classification scheme should not build on philosophies or theories but it should be derived from the published literature. Literary warrant or terminological warrant can be used in the same way for building thesauri. An index term is justified only if it occurs in the literature of the subject field often enough to be considered significant and useful in information retrieval. Literature sources where the terminology of the field is represented in a compact form are needed for deriving terms. Lancaster (1986, pp 25) considers abstract publications to be a good source for gathering terms.

In this study a variant of terminological warrant is used for developing a macro language. It is based on the indexing done by professional indexers. A macrolanguage reflecting the important concepts in a subject field can be developed on the basis of theories developed by Callon et al. (1983) and Law (1983). According to Callon et al. there are signal words which guide the readers selection of literature and describe the mechanism with which a scientific document attaches and keeps the interest of the reader. Law (1983) introduced the concept "interest funnel", which is organised around a few words, macro terms or signal words, syntetizing whole domains. The interest funnel is received through macroterms connected with other macro terms. The author of a documents use signal words to guide the readers in the direction he wants them to go and so takes the reader in to an interest funnel. The author has to use signal words accepted in the field. In content analysis the work proceeds in the other direction. The document is transformed to a string of signal words to catch its structure. To reduce a publication into a series of signal words does not mean that the ideas and information in the text are summarized. It means restoring of the texts function as an aid to guide the interest in a tunnel and define problems (Callon, 1983). These strings of signal words can also be used by the librarian for describing the information in the collections. How can librarians identify the signal words without deep subject knowledge of the field? In this study the expertise used by database producers for indexing is therefore used. An analysis of the indexing in the different databases shows the subject content of the articles according to the view of the professional indexer. It can be assumed that the persons indexing the documents have knowledge both about the subject field and about the terminology. On the other hand the indexer is influenced by his environment, for instance his own user group. Other factors which effect the indexing are the vocabulary or terminology used (the documentation language) and the depth and specificity of indexing. Because of this it seems important to study the indexing from the point of view of many indexers. In that way new strings of terms are developed based on a selection of terms which professional indexers have chosen to reflect the subject content of a specific article.

ISSN: 2324-9773

This was well reflected by the results of a project at Abo Akademi University where index terms were used for building a classification system. The aim of the project was to develop an expert system to aid selection of online databases (Trautman, 1988). A coded file of online database attributes was needed and the most challenging attribute was the subject classification of online databases. In the project subject indexes of standard online database directories were analysed. The editors of the directories, experts in database selection, have assigned subject terms (descriptors) to the online databases. Each of these terms represents the name of a family for which its members are already listed. The problem was how to present the subject terms to the user. The solution was to consider possible underlying goals, prejudices, interests, themes, or independent variables that the user might have when asking a question. All of these were called by the generic term "viewpoints." Ten first-level categories of viewpoints were adopted and have these brief names: TIME, FACTS, QUALITY OF LIFE, COMMUNICATIONS, COMMERCE, TECHNOLOGY, ADMINISTRATION, LIFE, ENVIRONMENT, and REALITY.

For each online database, the subject terms that had been assigned in each directory were noted. The original assignment, which was made in each of the directories, of individual databases to each term had been retained. This schedule was a composite index to those directories included and showed at a glance how the various editors differ in their assignment of subject terms to the online databases. The authors wanted to capture the various ways in which different human experts view the problem of database selection. The principle used in developing the three-level viewpoint classification was to integrate user information needs into generic categories. The goal was to facilitate new conceptual linkages in non-subject categories and interdisciplinary linkages in subject categories.

INDEXING CONSISTENCY

In information retrieval from bibliographic databases the representation of the content of the documents and the quality of that is essential. Each reference in a bibliographic database usually gives two types of information: the full bibliographic description and indicators of the subject content of the original text. The pårt of the reference which identifies the subject content varies between different databases, but includes usually some of the following fields: title, abstract, classification notation or category, index term (descriptor) selected from a thesaurus, free-text terms selected by an indexer or terms chosen from an authority list and an indication on special numeric or factual information in the original document. These variables usually influence the result of the information retrieval because most of them increase the precision. The use of a controlled vocabulary is the users most important aid to increase the precision. Since one of the goals of a controlled vocabulary is to gather all the documents indexed with a common term, this vocabulary also is important for the retrieval. A controllerad vocabulary indikates a certain consistency in indexing of a subject and consequently easy retrieval of information.

The process of indexing consists of finding essential concepts in a document and to translate these concepts into corresponding terms in the indexing language used. This process includes many subjective elements. Different indexers can consider different concepts as important, but also the selection of index terms can vary for the same concept. The terms *biocatalysis* and *enzyme* are for instance related associatively and can both be selected to describe an enzymatically catalysed chemical reaction. The terms *enzyme* and *cellulase* reflect same concept, but on a different level of specificity. There is an even greater difference between indexers work, if they use different indexing

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

languages. The consistency of indexing has often been analysed because of these variations. The consistency between indexers work is a quantitative measure of the degree to which two or more indexers agree on the terms reflecting the content of a document (Markey, 1984). Many studies show that there generally is a low consistency (e.g. Iivonen, 1989, Gerhard, 1993). The measure of consistency varies depending on the method of analysis. In one type of study of indexing in different systems terms are compared letter by letter and the use of different grammatical forms and synonyms makes the consistency lower (this applies to studies comparing use of different indexing languages). The consistency is lowest if used terms are compared, even if a controlled vocabulary is used, but it is higher when the indexed concepts are compared in stead of actual terms.

Indexing of a document is built up of terms which are justified in different ways (Iivonen, 1989). Core terms refer to the central concepts which different indexers usually index. The supplementary terms are on the other hand depending on the individual indexer, the environment and for whom the indexing is done. If the terms are chosen strictly on the basis of joint occurence, this could lead to that many supplementary terms are excluded and the core terms which represent the central concepts are the only one left. In that way the most important concepts of an interdisciplinary field could be recognized. When developing index terms for a specific collection of documents in an interdisciplinary field on the basis of indexing done by professional indexers, the consistency of the terms should in this way besecured by use of more than one set of index terms.

BIOTECHNOLOGY

The interdisciplinary field chosen for this study was biotechnology. It is a very old interdisciplinary field, which has developed fast and is changing continuously. The definition of the content varies and the researchers are connected through common methods, rather than a common knowledge base (von Ungern-Sternberg, 1994). A librarian developing local collections can therefore have difficulties in identifying the literature and organizing it. There are applications in a broad spectrum of other disciplines and subject fields. Biotechnology is based on natural sciences, medicine and technology, that is "hard" subject fields. Hard subject fields have in contrast to "soft" fields (e.g. humanistic) a well-defined terminology (Storer, 1968). The subject field is therefore well suited for development of a documentation language. In spite of the fact that the field has its knowledge base in many disciplines, there are relatively few publications in bibliographic databases, which claim to cover only biotechnology.

Reasons for the difficulty to organise information in biotechnology can be found in inconsistent definitions, a huge amount of information and finally in the weaknesses of the classification systems. Biotechnology is based on scientific research in biology, microbiology, biochemistry, biophysics, genetics, cellular biology, molecular biology, process engineering, industrial economy and other neighboroughing fields. Literature contains many varied definitions on biotechnology. Most of them agree on that certain processes utilizing biological organisms are biotechnology, but it is not always as clear which processes do not belong to biotechnology. There is still confusion about the terminology of the field. Literature of biotechnology, both in printed and in electronic form, is growing fast. There is a great number of new publications in the field and the results of research are published both in well established journals and in new ones. Besides the core, which is directly associated with biotechnology, relevant basic and applied research is found in the literature of many other disciplines. The most important classification systems, such as UDC and Dewey, have not regarded biotechnology as one field, but the subject is scattered to different fields

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY.
6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

in science, medicine and technology. This makes it difficult to identify the relevant material in the field.

THE STUDY

The material used for the present study was a collection of Finnish articles retrieved from four different international bibliographic databases, Pascal Biotechnologie (PAS), Life Sciences Collection-Biotechnology (LSC), Current Biotechnology Abstracts (CBA), and Derwent Biotechnology Abstracts (DBA).

Derwent Biotechnology Abstracts (DBA)

Derwent Biotechnology Abstracts (Derwent Publications, England) covers according to the producer research in all subfields of biotechnology. More than 1100 journals among other material are referred selectively. In Dialog Information Services the database covers material from 1982 and contained 140.948 references in Januari 1993. The empasis seems to be on commercial use of biotechnology. The index terms are based on a thesaurus (4 ed., 1992) and the terms are combined to descriptive phrases in the indexing. The indexing is complemented with terms which do not appear in the thesaurus.

Current Biotechnology Abstracts (CBA)

Current Biotechnology Abstracts (Royal Society of Chemistry, England) covers according to the producer all aspects of biotechnology. The material is selected from more than 900 journals and other documents. In Dialog Information Service the database covers material from 1983 and contained 49.964 references in 1993. Biotechnology is defined by the producer as development and use of biological systems and functions to get products useful for the humanity. The commercial use of biotechnology is empazised in the documentation. A controlled vocabulary is used for indexing and up to ten index terms per document. Both specific and general terms are used; for instance are all algae are indexed with genus-species names and all antibiotics with both the individual name and the term antibiotics.

Pascal Biotechnologie (PAS)

Pascal (Institut de l'Information Scientifique et Technique (INIST), CNRS, France). The database is multidisciplinary but contains a special section for biotechnology, Pascal Biotechnologie. This section covers accordig to the producer all subfields of biotechnology, but the commercial aspects are not as much empazised as in the above mentioned databases. A controlled vocabulary is used for indexing and both specific and general terms are used for an article.

Life Sciences Collection: Biotechnology Research Abstracts (LSC)

Life Sciences Abstracts (Cambridge Scientific Abstracts, USA) is a multidisciplinary database in the life sciences, but has a special section for biotechnology. According to the producer all aspects of biotechnology are covered, from basic research on laboratory level, with empasis on research in genetic engineering and applications in chemistry, pharmaceutics and agriculture, to process engineering and methods for production in large scale. The material covered is selected from 500

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

journals and other documents. A controlled vocabulary is used for indexing.

These four databases claim to cover all aspects of biotechnology, but definition of biotechnology varies. To get a core of articles considered by at least two producers as biotechnology a demand was made that an article had to be included in two databases to be selected. There were in total 909 references with Finnish corporate address 1987-1991 and 176 of them were common to at least two databases. These 176 references (referred to below as articles) formed the material in the study. All databases use a thesaurus for indexing material covered. The index terms (descriptors) for each article were retrieved from the databases. Only 160 articles of the 176 in total were indexed in at least two databases. 16 articles were indexed in just one database and the second database covering the same article contained no index terms for it. This was due to the fact that a special issue was indexed on the issue-level and not on article-level.

The number of index terms was counted for each article in each database. This created no problems for three of the databases. In the case of DBA the database producer uses a method where descriptors are combined with terms from natural language to create phrases. It was therefore important to make exact rules for how to count the index terms in this database. The following rules were used: Every descriptor in the thesaurus was counted as one term. Index terms based on free-text words and phrases, but not appearing in the thesaurus but which still gave new information about the content were counted as index terms. A few exceptions were made to this rule. Genus-species combinations were counted as one index term in spite of the fact that the thesaurus in most cases gave only genus. Enzyme codes taken from the "Enzyme Nomenclature" were not counted, only names of enzymes. Prepositions and other "small words" were not counted. The word "effect" was not counted as index term, because it usually did not give any unique information about the content. To make the process clearer we can take an example of indexing of an article covered by all four databases.

"Continuous production of lignin peroxidase by immobilized
Phanerochaete chrysosporium in a pilot scale bioreactor"

Pascal, Biotechnologie (PAS):
English Descriptors: Production; Continuous; Enzyme; Fermentation;
Phanerochaete chrysosporium; Entrapped microorganism; Bioreactor; Pilot plant; Polyurethane; Nylon; Performance evaluation; Optimization; Culture medium; Lignin peroxidase; Ligninolytic enzyme
Broad English Descriptors: Basidiomycetes; Fungi; Thallophyta

Life Science Collection, Biotechnology (LSC):
Descriptors: bioreactors; immobilized cells; lignin peroxidase;
production; Phanerochaete chrysosporium; yield

Current Biotechnology Abstracts (CBA):
DESCRIPTORS: Phanerochaete chrysosporium; cell, immobilized; fungi
CHEMICAL SUBSTANCE(S): lignin peroxidase

Derwent Biotechnology Abstracts (DBA):
DESCRIPTORS: lignin-peroxidase prep., Phanerochaete

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

chrysosporium immobilization nylon-web support enzyme
fungus

The article was indexed with 18 terms in PAS, with 6 terms in LSC, with 4 terms in CBA and with 8 terms in DBA. The number of index terms in DBA was counted in the following way: The thesaurus gives *Lignin peroxidase, Phanerochaete, Immobilization, Nylon, Support, Enzyme, Fungus* and *Prep.* as descriptors. In this case it was easy to count the number of index terms, because they were all listed in the thesaurus. In the next example it was more difficult to count the index terms.

Derwent Biotechnology Abstracts:

Descriptors: alpha-, beta-, gamma-cyclodextrin prep. from starch, Bacillus circulans var. alkalophilus cyclomaltodextrin-glucanotransferase,
effect of ethanol, dimethylsulfoxide * polysaccharide bacterium enzyme EC-2.4.1.19

Thesaurus terms (descriptors) were: Cyclodextrin,
Prep.,Starch, Bacillus, Cyclomaltodextrin-glucanotransferase, Ethanol, Polysaccharide, Bacterium, Enzyme

The term *Bacillus circulans var. alkalophilus* was counted as one term (Genus-species), *Alpha-, beta-, gamma-cyclodextrin* were counted as three terms and dimethylsulfoxide was considered to be an index term in spite of its absence from the thesaurus. The words *from* and *effect of* were not considered. The number of terms were 12.

Generally the results of the counting show that Pascal Biotechnologie and Derwent Biotechnology Abstracts in average contained about twice as many descriptors per article than Life Sciences Collection Biotechnology and Current Biotechnology Abstracts. White and Griffith (1987) have identified dimensions for the evaluation of indexing in online databases. **One of them is finely discriminating** among individual documents. This is a measure based on the average number of index terms (descriptors) assigned to each document in a collection (exhaustivity of indexing). The assumption is that each additional term expresses more of the content of the document and permits retrieval on a greater variety of conceptual dimensions. Indexing in the databases in biotechnology was thus according to this measure of higher quality in the former databases than in the latter. This theory can although be questioned. The number of index terms does not automatically reveal a deeper indexing and a quantitative measure of this kind has to be considered as indicative. A quantitative analysis which is not based on analysis of the concepts behind the terms includes a high risk of error. In one database an index term can be built up by an informative phrase (a precoordinated term), in another database the parts of thew same phrase can be simple terms (postcoordinated terms) and the information content is then not compatible.

When the index terms were counted for each article new strings of index terms were developed based on the indexing in the databases. Sievert and Verbeck (1987) have made rules for converting terms to concepts.: 1. The terms are different grammatical or syntactical forms of the same concept (e.g. *information storage and retrieval* and *information retrieval*), 2. The term is a subordinated term to a more general term and there is a word in common (e.g. *reference services* and *library services*), 3. The terms are synonyms.

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

In this study these rules were applied in the following way: The index terms from the databases were sorted alphabetically for each article and the new string of terms was created through a selection of terms in common for at least two databases. If an index term was exactly in the same form it was included immediately. Singular forms were converted to plural if praxis varied in the databases, and capitals were changed to small letters. Inverse wording was changed into direct and abbreviations were written out if the same term was written out in another database; e.g. *ELISA* was changed to *Enzyme-linked immunosorbent assay*. Quasisynonyms were standardised through selection of one of the terms. Such quasisynonyms were for instance *Preparation - Production*, and *Biodegradation - Degradation*. Precoordinated terms were split up if they occured only in one database, but the separate terms occured in another database. So the term *gene* was chosen when *Gene manipulation* occured in one database and *Gene transformation* in another. This was to secure that the concept gene was included in the new string of index terms. The string of terms for the article in the example above was:

"Continuous production of lignin peroxidase by immobilized Phanerochaete chrysosporium in a pilot scale bioreactor"

Index terms	Occurence in number of databases
Lignin peroxide	4
Phanerochaete chrysosporium	4
Immobilized cells	3
Production	3
Bioreactor	2
Enzyme	2
Fungi	2
Nylon	2

The term *Production* occured in two databases och the DBA term *Prep.* was according to the rules changed to *Production*. The inverse wording in CBA: *Cells, immobilized* was changed into a direct. The result was 8 index terms. The four databases contained totally 36 terms for this article, that was in average 9 per database. 8 new terms in the new string corresponded well to this average number.

The number of index terms in the new strings show that at least three databases are needed as base for developing terms. This indicates that the consistency between indexers is low. The enhancement of the strings of new index terms based on titles shows the same. Only one of the 40 articles indexed in all four databases and two of the 54 articles indexed in three databases were complemented. On the other hand the strings of index terms for 21 of the 66 articles covered by two databases had to be complemented. This method of developing terms on the base of indexing in different databases works thus only if at least three databases are used.

The new string of index terms for each article was checked with the title to ensure that no important aspect of the content was excluded. If this was the case the string of terms was complemented with an additional index term, which occured both in the title of the article **and** as an index term in one database. Such complemented strings of index terms had to be made for 25 articles (14,2%). The reason for doing this was to ensure that terms which described the same concept, but were given as

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

different terms in the databases were identified. The 25 articles were scattered to the databases in the following way: One of them was covered by all four databases, three articles were covered by three databases and 21 were covered by two of the databases. Almost 32% of the new strings of index terms for articles indexed in two databases had so to be complemented with an additional index term. The following example can demonstrate the process of complementation:

Monoclonal antibodies to the 27-34k **insulin-like growth-factor**
binding-protein.

LSC: **insulin-like growth factor**; binding; proteins; monoclonal antibodies; characterization; placenta (6 index terms)

DBA: somatomedin binding protein monoclonal antibody prep.,
characterization, hybridoma construction * mammal cell culture
(10 index terms)

The index terms in the new string are: *Binding, Monoclonal antibody, Protein*. In the new string of index terms the growth factor mentioned in the title is lacking. In DBA there is a term *Somatomedin*, which describes a growth hormone and in LSC there is the term *Insulin-like growth factor*. The new string of terms is complemented with the term *Insulin-like growth factor*.

The new index terms were sorted alphabetically and standardized according to different ways of spelling and grammatical forms. The strings of new index terms contained in total 427 different terms and 255 of them occured just once.

A discrimination index was calculated for the 427 terms according to Ajiferukes (1988) method: Discrimination index (term A)= n (term A)/N, where n=the number of documents which are indexed with the term A and N=total number of documents in the collection. This index is also between 0 and 1. The lower the index, the more discriminating term. Index is 1 if the term has been used for indexing of every document in the collection and 0 if it is not used. Good discriminating terms have values between 0,001 and 0,05.

According to this the 22 most frequent terms had a discrimination index of at least 0,05. The 22 most frequent terms can be considered to be macroterms, which generally reflect the subject content of the collection.

ISSN: 2324-9773

Frequency	Index term	Discrimination index
57	Enzyme	0,356
46	Fungi	0,288
44	Bacteria	0,275
38	Gene	0,238
29	Production	0,181
25	Trichoderma reesei	0,156
23	Cloning	0,144
20	Plasmid	0,125
18	Alpha-amylase	0,113
17	Bacillus subtilis	0,106
16	Escherichia coli	0,100
15	Saccharomyces Cerevisae	0,094
14	Yeast	0,088
13	Gene expression	0,082
12	Characterization	0,075
12	Purification	0,075
10	Cellulase	0,063
10	Fermentation	0,063
10	Immobilization	0,063
10	Vector	0,063
9	Degradation	0,056
9	Expression	0,056

Table 1. The most frequent terms and their discrimination indexes.

The most frequent terms included general terms such as *Enzyme, Fungi, Bacteria* and *Gene*, but also very specific bacteria used in genetic engineering and microbiology. Nine of the 22 most frequent terms described methods or techniques. *Trichoderma reesei, Bacillus subtilis, Escherichia coli* and *Saccharomyces cerevisae* are the best known model organisms in genetic engineering. They are modified genetically e.g. to produce enzymes which better can stand different circumstances (e.g. enzymes in detergents). The alpha amylase gene can be cloned and the gene is transferred to a bacteria as *Bacillus subtilis*, which then will produce the enzyme.

The 22 terms were distributed to three main subject fields:
Microbiology: *Enzyme, Fungi, Bacteria, Yeast, Fermentation, Immobilization,*
Genetic engineering: *Gene, Cloning, Plasmid, Gene expression, Vector, expression.* **Methods**: *Production, Characterization, Purification, Degradation.* Genetic engineering can on the base of this be considered as core field in biotechnology. An analysis of all the 427 terms according to their content gave the following distribution to different categories:

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

Micro organisms (bacteria, viruses,yeast a.s.o)	63
Methods, production technology	60
Enzymes	48
Chemical compounds	44
Genetic engineering, genetics	40
Plants and parts of plants	36
Physical phenomenon and properties	20
Cell- and tissue cultures	19
Cells, immunological terms	11
Proteins, nucleic acids, peptides	11
Waste	9
Wood, paper, cellulosa	9
Food	7
Sugar, starch	6
Others, general terms	44

Table 2. Categories of index terms.

The largest categories of terms were microorganisms, methods, enzymes, chemical substances and terms in genetic engineering. There were also many terms describing methods and production technology. Production of enzymes and biocatalysis was a strongly emphazised group among the terms, as also genetic engineering, technological terms and botanical terms (plants and parts of plants). There were surprisingly few terms in medicine.

Trofimenkos method can be applied to the terms developed in biotechnology and gives an indication about the state of the research in Finland. The frequencies of the terms indicate that the field is still developing and that there is a high degree of interim research. No term has high frequency if the level is set at half of the articles (Ajiferuke, 1988, White, 1987). The high frequent terms reflect according to Trofimenko the stable and most developed directions of research. The three most frequent terms are micro organisms (*Enzyme, Fungi, Bacteria*) and could then be considered as representing the stable part of biotechnology. There is about 60% low frequent terms. In Trofimenkos investigation the part of low frequent terms in different fields varied between 33% and 64%. The field *Nuclear Power Plants*, which also was the largest field in Trofimenkos investigation, had the lowest number of low frequent terms in relation to all the used terms. Compared to Trofimenkos study biotechnology can be considered as a field in development with a high volume of occasional research. This can be partly be explained through the fact that the commercial development of different products is high in biotechnology and there is a search for new applications.

METHODOLOGICAL PROBLEMS

There were many problems in developing the strings of new terms. Different databases use different depth and specificity for their indexing. The way in which issues of a journal addressing a special topic are handled varies; they may be considered as one collective reference or many individual articles. There is obviously a need for a professional standard which would show the indexing policy in the database documentation. This standard ought to tell how exhaustive the indexing is, which kind of articles are indexed and which not, who makes the decisions and on what base. The

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY.
6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

criteria for selection of documents for inclusion in the databases should also be explained in the manuals. Indexing is rather primitive in most secondary information sourcęs (Rader, 1988). The subject indexes use either very simple and general classification systems, use index terms without vocabulary control or utilize the indexing language of information sources with broader subject coverage. For biotechnological information the classification systems and vocabularies are weak. The lack of relevant classification systems and vocabularies has a negative influence on the whole infrastructure of information systems by making information retrieval and coordination of resources more difficult and the organization and exchange of information on a very simple basic level.

The articles indexed in the same way in each of the four databases have a higher impact on a consistent documentation language than the articles for which the indexing varies much in the different databases. A weigthing of the terms could be used according to their oçcurence in the databases.

SUMMARY

Collection development does not end in selecting relevant material. When literature is acquired to a library collection its content has to be analysed systematically and the information resulting from this analysis has to be organized and made available. A documentation language is needed. In this study an attempt has been made to develop terms for an indexing language reflecting Finnish research in biotechnology and to define the information content of the field. The terms have still to be tested through use and semantic relations have to be developed (e.g. ISO 2788) before they can be a basis for an indexing language. The method could be used for new subject fields, which are not well defined, to get an aid for analysis of the information content of local research and for identifying and organizing the literature in the field. The terms are based on work done by professional indexers for international, bibliographic databases. These terms reflect new, published research and are based on a collection of articles selected by several database producers, who claim to cover all aspects of biotechnology. Since there has been two to four strings of index terms for each article and the new strings of index terms have been selected on the basis of these, they can be considered to give an objective view of the field.

If the method is applied to a subject field without bibliographic databases, a core collection of relevant articles could be gathered and index terms could be retrieved for them from bibliographic databases covering the journals they are included in. If the same articles for instance are included in databases covering different disciplines as Chemical Abstracts, Biosis and Medline, the common indexing could give an objective view of the information content and a base, which can be used for retrieval of more relevant articles and their index terms. The method developed has two functions: It will give the librarian a local basis for defining a new subject field and it will give terms, which can be used for building a controlled vocabulary reflecting the new research in the field.

LITERATURE

Ajiferuke, I., Chu, C. M. (1988). Quality of indexing in online databases. An alternative measure for a term discriminating index. *Information Processing & Management* 24, 599-601.

Beghtol, C. (1986). Semantic validity: concepts of warrant in bibliographic classification systems.

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY. 6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

Library Resources & Technical Services 30, 109-125.

Callon, M., Courtial, J.-P., Turner, W. A., Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* 22, 191-235.

Gerhard, K. H., Jacobson, Trudi E., Williamson, S. G. (1993). Indexing adequacy and interdisciplinary journals: the case of women's studies. *College & Research Libraries* 54, 125-133.

Iivonen, M. *Indeksointituloksen riippuvuus indeksointiympäristöstä.* [Indexing and the indexing environment.] (1989). Tampere: Tampereen yliopisto. Tampereen yliopiston kirjastotieteen ja informatiikan laitoksen
tutkimuksia. 26. (in Finnish).

ISO 2788. Guidelines for the establishment and development of monolingual thesauri. International Organization for Standardization, 1974

Law, J. (1983). Enrolement et contre-enrolement: Les luttes pour la publication d'un article scientifique. *Social Science Information* 22, 237-251.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval.* 2. ed. Arlington: Information Resources Press.

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, 155-177.

Rader, R. A. (1988). Status of the infrastructure of information resources supporting US biotechnology. In: *The impact of chemistry on biotechnology.* Multidisciplinary discussions. Ed. by Marshall Phillips etc. Washington, DC: American Chemical Society. ACS Symposium Series pp. 375-385.

Sievert, M. E., Verbeck, A. (1987). The indexing of the literature of online searchin: a comparison of ERIC and LISA. *Online Review* 11, 95-104.

Storer, N. W. (1967). The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association* 55, 75-84.

Trofimenko, A.P. (1990). Scientometric analysis of the topical content of scientific research and its particularities. *Scientometrics* 18, 409-435.

Trautman, R., von Flittner, S. (1989). An expert system for microcomputers to aid selection of online databases. *The Reference Librarian* (23), 207-238.

von Ungern-Sternberg, S. (1994). *Verktyg för planering av tvärvetenskaplig informationsförsörjning.* En tillämpning på ämnesområdet bioteknik i Finland. [A tool for planning information provision in interdisciplinary fields, applied to the field biotechnology in Finland.] Abo: Abo Akademis förlag. (in Swedish).

ISSN: 2324-9773

Von Urgern-Sternberg, S. (1995). KNOWLEDGE ORGANISATION AND A MACRO LANGUAGE, FOR BIOTECHNOLOGY.
6th ASIS SIG/CR Classification Research Workshop, 193-206. doi:10.7152/acro.v6i1.12669

White, H. D., Griffith, B. C. (1987). Quality of indexing in online data bases. *Information Processing & Management* 23, 211-224.

ISSN: 2324-9773