Data structuring and classification in newly emerging scientific fields

(An information method based on bibliometric analysis of keywords used for data structuring and classification in newly emerging scientific fields. Example: functional food)

Simona JUVAN
Information specialist. Central Biotechnical Library, Biotechnical Faculty, University of Ljubljana, Slovenia.

Tomaz BARTOL
Assistant Professor of Information Science at the Department of Agronomy, Biotechnical Faculty, University of Ljubljana, Slovenia.

Bojana BOH
Professor of Scientific and Technological Informatics at the Department of Chemical Education and Informatics, Faculty of Natural Science and Engineering, University of Ljubljana, Slovenia.

Authors' version of Research paper
Published in: ONLINE INFORMATION REVIEW, 2005, vol. 29, no. 5, p. 483-498

Keywords
bibliometric analysis, bibliographic databases, data processing, indexing, subject headings, thesaurus, thesauri, data structures, keyword analysis, structuring data into systems, classification, functional food, foods, human nutrition, bibliometrics, scientometrics, data mining

Abstract
The article addresses a methodological procedure based on keyword analysis and structuring of data into information systems in the field of functional foods as a newly emerging scientific field within the broader scope of food sciences and technology. Experiment was based on selection of a research field or research subject, selection of search profile, selection and processing of relevant databases, keyword analysis, and arranging of data (keywords) according to tree-structures. Keyword analysis was employed to identify narrower research fields within the broader scientific field. Structuring of data into systems was used to systematically classify the terms within the particular narrow field. Keywords with higher and lower frequency were identified. A classification tree was set up, based on keywords (thesaurus-based descriptors) extracted from FSTA database available online. The tree was supplemented and upgraded with some additional topical terms, that have not as yet been included in the existing thesaurus. To serve as a comparison a completely new classification tree was designed based on online full-text data. The comparison of the two trees suggests the previous existing tree as being sufficiently accurate to represent the field of functional foods, providing that it be upgraded with some additional terms. A more accurate classification should improve thesauri and consequently enhance retrieval in international databases.

INTRODUCTION

Rapid growth of information and availability of data from an increasing amount of new sources have boosted possibilities of data and information exploration, and, consequently, of identification of research trends and patterns in a particular area of knowledge. In the field of library and information science, quantitative studies of information have been marked by evolution of bibliometrics (Sanz-Casado et al., 2002). Bibliometric analyses can thus be very useful in following research trends (Mela et al., 1999).

In most bibliographic or full text databases, users can search either within free text, or with controlled terms (keywords or descriptors). Controlled terms allow a searcher an improved retrieval precision of documents on a selected topic, regardless of the words that the authors of a particular document had assigned to the topic. Indexing, however, generally depends on human analysis of the subject. Indexers read the material, locate the best terms in a thesaurus, and assign the terms that describe a document’s meaning. Common types of subject description in databases are usually arrayed from broad to specific.
Studies of indexing show significant variation in the use of terms chosen by different indexers to represent the same topic or document (Bertrand et al., 1995). Suraud et al. (1995) observed non-existence of well-defined keywords in newly emerging fields, what makes bibliographic searches difficult. Bates et al. (1993) call for development in the structures of thesauri and in design of online information systems. Hurt (1997) maintains that it is necessary to renew and expand indexing and classification systems. Soergel et al. (2004) also point out that existing classification schemes and thesauri lack in well defined semantics and structural consistency.

The above indicates that the existing systems do not necessarily allow for sufficiently accurate and correct information retrieval so some further development and improvement of search systems might be useful. Newly emerging scientific fields in particular need a more correct classification and better inclusion of pertaining terms in international bibliographic thesauri.

In our research we will try to characterise a newly emerging scientific field (functional foods) with regard to some principles of database indexing. We will identify and assess the existing indexing terms and existing classification tree in leading bibliographic databases in the field. We will complement and thus upgrade, where necessary, the existing indexing terms (descriptors) with some additional suitable terms which have hitherto not been assigned to the documents or not as yet been proposed as indexing terms in the existing thesauri. Also, we will analyse some selected online full-text documents in order to create a completely new classification tree. We will compare the existing, the upgraded, and the newly proposed classification trees.

The following methodological approach is offered as contribution to facilitate and improve the process of creating classification structures and thesauri in newly emerging scientific fields, and consequently to improve database quality and retrieval of documents.

**MATERIALS AND METHODOLOGY**

Our methodology was based on three concepts: (1) an information method to achieve value-added processing of bibliographic databases in science and technology (Kardos et al., 2000), complemented by (2) a bibliometric analysis (Mela et al., 1999) to define research trends and obtain keywords, and (3) a method to structure the data into systems (Kornhauser, 1989) for the classification of keywords.

The procedure was performed through the following steps (Figure 1): (1) selection of a research field, (2) definition of a search profile – search query, (3) selection and processing of an appropriate database, (4) extraction and analysis of keywords from the records retrieved with a well defined search query, and (5) arrangement of keywords into a tree-structured classification system.
Selection of a research field

As the first step a specific "research field of interest" was defined and the relevant scientific terms were identified.

Definition of a search profile

After the research field had been selected, an appropriate search profile or query was set up, and the time scope for the retrospective search and retrieval was defined. The search query typically consists of a combination of keywords and a time series, and some other possible searchable parameters.

In the case of a relatively new scientific field, the search query needs to include all potential synonyms, and needs to be based on all entries, not only on the descriptor entry, to minimise the risk of loosing documents on account of potential terminological inconsistencies in newly developing concepts.

This is in accordance with the observation of Suraud et al. (1995) whereby uncontrolled entries like title, abstract and free terms better reflect evolution of a concept and might be more easily adaptable to an emerging concept, as opposed to descriptors and classification code entries, which offer a relatively rigid access to a document.

Selection and processing of appropriate databases

In the third step, the most relevant database was selected. Within a group of potentially interesting databases, selection can be conducted on the basis of a search preview tool or
similar search options. In addition to the number of relevant records, other criteria can be considered (Kardos et al., 2000), such as database content, availability, and price.

**Extraction and analysis of keywords**

In the fourth step, keywords were extracted from the retrieved records, and were classified into groups according to main research trends within the scope of the selected research topic. This classification is best made on the basis of *title* and *abstract* of a record.

In our keyword analysis, a method by Mela et al. (1999) was used, where keywords were defined as comma-separated items of one or more words. Different keywords with identical meaning (synonyms) were identified "manually", and were further assigned a single-keyword-meaning. The keywords were then arranged into sub-groups of related terms.

**Structuring of keywords into a tree-structured classification system**

In the fifth step, arranging of keywords according to tree structures was carried out in the following order: (1) arranging of parameters into large sets of data - in our case an identification of parameters whereby keywords can be classified into sub-groups (Kornhauser, 1989); (2) defining hierarchy of parameters - arranging of keywords according to a broader or narrower (more specific) meaning, and (3) arranging of data into a tree-structured system, i.e. developing a logical classification system, which includes all extracted keywords.

The method allowed us to build a logical classification system according to the standard principle of arranging the terms from a broad concept to more specific, narrower concepts.

As a research material and as a terminological reference we used online bibliographic databases, Web pages, full-text journals by major scientific providers, and books in libraries.

**RESULTS AND DISCUSSION**

In the case of our research a method itself can also be interpreted as a part of the results of the analysis. We thus present some of the methodological issues also in the following field.

**Selection of a research field**

The concept of *functional foods* was selected as an example of a newly emerging scientific field. According to an accepted definition, functional foods are foods which are consumed as a part of normal daily diet, and which have health promoting benefits and reduce risk of diseases (Diplock et al., 2000; Roberfroid, 2000).

**Definition of a search profile – search query**

The concept of *functional foods* is not identical to the notions of *medical foods* or *dietary supplements*. It overlaps, however, with the notions of *foods for special dietary uses* and *fortified foods*. Functional foods may in some cases also be expressed as *nutraceuticals*, *health foods* or *novel foods* (Kwak et al., 2001).

Our search profile involved the 1990 – 2002 period, and entailed the following terms: functional food* or nutraceutical food* (employing right truncation). The retrieval with these terms was conducted in abstracts (AB), document titles (TI), and database keywords or descriptors (DE). Further examination, however, involved only the extraction and analysis of keywords from the Descriptor field.

**Selection and processing of databases**

The truncated search query functional food* was employed to identify databases with the highest number of records in order to select the most appropriate database for further analysis.

The growth of publications on functional foods in 1990-2003 is presented in Figure 2. In most databases a rapid growth of documents can be noticed after 1997, what corresponds with the general recognition and acceptance of the definition of the term functional food. After 2001 the numbers of new publications show signs of some decrease. This may perhaps be explained by the fact that the research field functional food has now somehow stabilised what is mirrored also by the number of newly published documents.

![Figure 2: STN Free Search Preview – growth of records in bibliographic databases that cover food science and technology research; search profile: functional food* in the period 1990-2003; abbreviations are explained in Table 1 (processed 7.1.2005)](image-url)
Table 1: Examples of bibliographic databases containing data relevant to the functional food research (selected from the STN International host – Food Science and Technology Cluster, processed 21.4.2004)

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Database Type</th>
<th>Coverage</th>
<th>File Size April 2004</th>
<th>Producer</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>CABA</td>
<td>Bibliographic database</td>
<td>1973 to the present</td>
<td>More than 4,500,000 records</td>
<td>CABI Publishing, UK</td>
<td>Agriculture and related science including veterinary medicine, human nutrition, developing countries, leisure, recreation, and tourism.</td>
</tr>
<tr>
<td>CAPLUS</td>
<td>Bibliographic database</td>
<td>1907 to the present</td>
<td>More than 23,590,000 records</td>
<td>Chemical Abstracts Service, USA</td>
<td>All areas of chemistry and chemical engineering. The major subdivisions are: Applied Chemistry and Chemical Engineering, Biochemistry, Macromolecular Chemistry, Organic Chemistry, Physical, Inorganic, and Analytical Chemistry.</td>
</tr>
<tr>
<td>FOMAD</td>
<td>Bibliographic database</td>
<td>1982 to the present</td>
<td>More than 181,000 records</td>
<td>FIZ Karlsruhe, Germany</td>
<td>Detailed analyses of international food and drink markets, identifying key market players, highlighting new product launches, assessing consumer attitudes and retail trends, and tracking company news.</td>
</tr>
<tr>
<td>FROSTI</td>
<td>Bibliographic database</td>
<td>1972 to the present</td>
<td>More than 549,000 records</td>
<td>The Leatherhead Food Research Association, UK</td>
<td>All aspects of the food and drink industry are covered.</td>
</tr>
<tr>
<td>FSTA</td>
<td>Bibliographic database</td>
<td>1969 to the present</td>
<td>More than 600,000 records</td>
<td>IFIS Publishing, UK</td>
<td>All areas which relate to food and beverages including the latest research, new product development, quality assurance and control, food safety, food composition, regulations, consumer trends.</td>
</tr>
<tr>
<td>JICST-EPlus File</td>
<td>Bibliographic database</td>
<td>1985 to the present</td>
<td>More than 4,723,000 records</td>
<td>Japan Science and Technology Agency (JST)</td>
<td>The literature published in Japan on all fields of science, technology, and medicine.</td>
</tr>
<tr>
<td>PROMT</td>
<td>Multiple-industry database</td>
<td>1978 to the present</td>
<td>More than 11,234,000 records</td>
<td>Gale Group, USA</td>
<td>International coverage of companies, products, markets, and applied technologies for approximately 60 manufacturing and service industries.</td>
</tr>
</tbody>
</table>

Explanation of database acronyms:
CABA: CAB Abstracts
CAPLUS: Chemical Abstracts Plus Database
FOMAD: Foodline®: Market Sight
FROSTI: Foodline®: Science Sight
FSTA: Food Science and Technology Abstracts
JICST-EPlus File: JICST file on Science, Technology, and Medicine in Japan
PROMT: Predicasts Overview of Markets and Technology
Other international bibliographic databases, which also include records on food science and technology research (e.g. AGRIS and MEDLINE) contained fewer than 1000 functional-foods-related records in the observed period, and were consequently excluded from further analysis.

FSTA database was selected for further analysis as the principal source of relevant records. The 1990-2002 period query "functional food* or nutraceutical food*" retrieved 1555 different records (1164 journal articles, 133 reviews, 79 conference proceedings, 77 patents, and 102 other type documents such as thesis, books, reports). A rapid increase of journal articles can be especially noticed in the more recent period (Figure 3).

![FSTA: functional food* or nutraceutical food*/DT (1990-2002)](image)

**Figure 3:** FSTA: yearly growth of records by document type, search profile: functional food* or nutraceutical food* (1990-2002).


**EXTRACTION OF THE EXISTING KEYWORDS**

**Keyword analysis**

Our bibliometric examination harvested 1555 initial records. The total corpus of keywords was 8367, including multiple occurrences of the same keyword in different records. The total of different original database-assigned keywords was 1223. Prior to the examination we proposed, according to our expertise in the field, the following possible topical groups: **foods, classification** (of biologically active compounds), **properties** (of biologically active compounds), **physiological effects**, **foods** (in general sense), **diseases, legislation, and processes**.
Group Foods represents those documents where a food (a fruit, vegetable, meat product etc.) was referred to in general sense. Group Classification contains those keywords which represent a specific biologically active compound, such as calcium, inulin, lactoferrin, folic acid. Group Properties represents some physico-chemical and organoleptic properties (e.g. physical properties, colour, solubility, texture) of foods. Our topical clustering of keywords and assessment of documents also yielded some other documents with the prevalence of documents described as bibliography which were usually some type of indexes, and which we excluded from topical subject analysis. The groups also presented in Table 2.

Table 2: Keyword structuring of functional foods into main topical groups with regard to number of different keywords, and database keyword occurrence.

<table>
<thead>
<tr>
<th>Group</th>
<th>Different terms (keywords)</th>
<th>Keyword occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOODS (e.g. yoghurt, eggs, chocolate)</td>
<td>419</td>
<td>3077</td>
</tr>
<tr>
<td>CLASSIFICATION (e.g. calcium, inulin, lactoferrin, folic acid)</td>
<td>325</td>
<td>1761</td>
</tr>
<tr>
<td>LEGISLATION (e.g. regulations, policy, standards)</td>
<td>124</td>
<td>1128</td>
</tr>
<tr>
<td>DISEASES (e.g. diabetes, osteoporosis, colon cancer)</td>
<td>79</td>
<td>869</td>
</tr>
<tr>
<td>PROCESSES (e.g. chromatography, encapsulation, freezing)</td>
<td>179</td>
<td>767</td>
</tr>
<tr>
<td>BIBLIOGRAPHY (e.g. patents, reports)</td>
<td>10</td>
<td>375</td>
</tr>
<tr>
<td>PHYSIOLOGICAL EFFECTS (e.g. immunology, digestibility, antitumour activity)</td>
<td>41</td>
<td>233</td>
</tr>
<tr>
<td>PROPERTIES (e.g. physical properties, colour, solubility, texture)</td>
<td>36</td>
<td>129</td>
</tr>
<tr>
<td>OTHER</td>
<td>10</td>
<td>28</td>
</tr>
</tbody>
</table>

Frequency or occurrence was calculated for each keyword, and was specified as a parameter of information density. The term information density is characterised as a frequency of occurrence of a separate search term in a separate search field (Boh, 1996; Kardos and Boh, 2000), and can be expressed as a percentage (e.g. 20%, if a search term appears in 20% of records), or as a number.

**Structuring of keywords into a classification tree system**

Our further analysis involved only the keywords from the above group that was tagged as Classification. Namely, this group comprised only the biologically active compounds, e.g. calcium, beta carotene, inulin, lactoferrin etc. Because of the existing inconsistent chemical classification of biologically active compounds in functional foods the method of structuring of data into systems (Kornhauser, 1989) was used to propose an improved and more consistent classification tree.

In most cases we used the FSTA Thesaurus (2004) as a term of reference. In some cases, however, this thesaurus was not suitable because it is not always in accordance with the standard chemical classification of active compounds.

An example of insufficiency of this thesaurus is the arrangement of carbohydrates which are generally a well established chemical group. Cummings et al. (1997) divide carbohydrates into monosaccharides (glucose and fructose), disaccharides (sucrose, lactose, etc.), sugar alcohols (sorbitol, maltitol, etc.), oligosaccharides (maltooligosaccharides and nondigestible oligosaccharides) and polysaccharides (starch and non starch polysaccharides). FSTA Thesaurus (2004), on the other hand, divides carbohydrates into bifidus factors, glycosides,
oligouronides, polyols, polysaccharides and sugars (amino sugars, disaccharides, oligosaccharides, reducing sugars, etc.). In this case we gave priority to the afore-mentioned classification by Cummings to obtain a more accurate chemical classification and to avoid having too many hierarchical levels.

UPGRADING OF THE EXISTING KEYWORD STRUCTURES

During the process of data structuring, some keywords were excluded, such as keywords that did not represent specific compounds (e.g. keywords ingredients, composition) or some general group that could not be placed into a specific chemical group. Those keywords that generally stood for a functional group we arranged into a corresponding chemical group.

We identified synonyms (e.g. vitamin C / ascorbic acid) and then merged these synonyms as a single term. We structured the remaining keywords from the group Classification into subgroups according to the chemical nature of a particular compound. We thus created additional two main "chemical" branches of the tree system Inorganic compounds and Organic compounds. We added these two groups to the already existing field of Microorganisms (Figure 4). Figure 4 represents the left fraction (main branches) of the Figure 5. White fields represent the existing thesauri-derived keywords with high information density. Grey fields in the figures represent additional categories which we created in order to better describe and explain chemical structure of the field of functional foods.

![Classification Tree](image-url)

Figure 4: Main branches of classification tree of biologically active compounds in functional foods, based on existing (white areas) and supplemented (grey areas) terms.

In some cases we thus created a new category (according to chemical classification) that had not existed in the original classification structure or the initial set of keywords and which had...
not been obtained during the bibliometric analysis. We defined such a group with the help of a thesaurus or some other source of classification of chemical compounds. This consequently resulted in a more consistent classification in our experimental database.

While counting keywords we observed that those with higher frequency (higher information density) are usually represented as main branches on the left side of the classification tree (e.g. minerals, vitamins, antioxidants), while keywords with a lower frequency, such as individual biologically active compounds (e.g. calcium, vitamin A, quercetin, zeaxantin), represent the branches of lower hierarchical order on the right side (Figure 5).

We noticed, however, a few exceptions. In some cases high-frequency keywords come about in lower branches, such as the keyword inulin with 24 occurrences, and which had been assigned the position of the sixth branching (the furthest from the root node in the Classification). Most other keywords with similar occurrences were placed in the second or third branching of the tree. Another example of an exception is the keyword fructans with four occurrences. It was positioned in a branch closer to the root node than inulin even though fructans are a broader concept than inulin (according to the FSTA Thesaurus) and should occur more frequently.

The analysis and structuring of keywords indicate that the indexing of the FSTA database frequently differs from the standard chemical classification. Some records which contain keywords glucose and fructose represent an example of inconsistence. Both substances can be chemically classified as monosaccharides. However, this term (monosaccharides) was not assigned as a descriptor to the respective records in the FSTA database, even though the term monosaccharides does feature as a descriptor in the FSTA thesaurus. In some cases, but again not consistently, there occurred the keyword sugars. From the chemical point of view, where carbohydrates are usually divided into monosaccharides, disaccharides, oligosaccharides and polysaccharides, for the keywords glucose and fructose the group term monosaccharides would be more appropriate than the term sugars (as sugars can be monosaccharides or disaccharides) or at least it would be better to use both keywords: sugars and monosaccharides. This leads to conclusion that indexing of this database differs from some principles of the standard chemical classification and that the broader terms are in certain cases used less systematically.

Theses findings concur with some previous research. For instance, Cohen et al. (1979) noticed that relations between broader and narrow terms in the FSTA thesaurus were not always accurate. Weintraub (1992) compared searches in agricultural databases such as AGRICOLA, AGRIS and CAB with regard to the topic of alternative agriculture and established that standardised terminology was lacking. Hutchinson et al. (1996) pointed out lack of standardisation in spelling and of certain terminology as the main weakness of FSTA thesaurus. Also, Bartol et al. (2000) stressed that in some databases, such as AGRIS, CAB, FSTA and MEDLINE, many different terms exist for the same topic, and these can sometimes be used as a narrow and sometimes as broader term without inter-database consistency.

One of the methodological problems in structuring of keywords into a classification tree is to determine which keywords of higher (broader) information density can substitute some other more specific (narrower) keywords. We tested the above by extracting and structuring only the keywords with the frequency of four or more occurrences from the group of biologically active compounds. The experiment indicated that the significant groups of biologically active compounds (e.g. vitamins, minerals, antioxidants) and important individual compounds (e.g.
calcium, vitamin E, lactoferrin) are still present in the tree system as shown in Figure 4. This result suggests that it might be sufficient to use only the keywords with higher information density, however, with some restriction. The usage of only higher-information-density keywords will capture most broader chemical categories. On the other hand, some documents may be lost, especially those describing novel compounds and not as yet clearly defined new scientific terms.

An example are the organic sulphur compounds found in garlic. Garlic is a functional food with more than 30 organic sulphur-containing compounds, such as alliin as a biologically active compound (Markowitz et al., 2003). Our study found out that there was only one keyword representing the organic sulphur compounds (isothiocianates with 2 occurrences). Alliin was not present among the keywords, even though alliin does come about as a descriptor in the FSTA Thesaurus, as showed another search in FSTA database (Table 3) with different search profiles containing keywords garlic, alliin and functional food*. Records that contained the term alliin had no correlation to the term functional food. An analysis based strictly on keywords might conclude that organic sulphur compounds have no significance as biologically active compounds. Such a conclusion would be incorrect because it would be based on deficient indexing of documents due to inadequate knowledge on the part of the indexing staff. The results therefore indicate that in certain cases indexing in FSTA database is not conducted in comprehensive agreement with the principles of chemical classification. Such deficiencies are not a characteristic of FSTA thesaurus only. Most thesauri exhibit some level of shortage or subjectivity with regard to terminological representation. Moreover, this fact applies not only to new emerging scientific research fields but also to basic fields, e.g. chemistry in general.

Table 3: Documents in FSTA database (1990-2002) containing different combinations of keywords garlic, alliin, and functional food*.

<table>
<thead>
<tr>
<th>Search profile</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>alliin</td>
<td>63</td>
</tr>
<tr>
<td>alliin in de</td>
<td>42</td>
</tr>
<tr>
<td>garlic</td>
<td>1013</td>
</tr>
<tr>
<td>garlic in de</td>
<td>483</td>
</tr>
<tr>
<td>alliin and garlic</td>
<td>47</td>
</tr>
<tr>
<td>(alliin and garlic) in de</td>
<td>19</td>
</tr>
<tr>
<td>alliin and functional food*</td>
<td>0</td>
</tr>
<tr>
<td>garlic and functional food*</td>
<td>16</td>
</tr>
<tr>
<td>(alliin in de) and functional food*</td>
<td>0</td>
</tr>
<tr>
<td>(garlic in de) and functional food*</td>
<td>1</td>
</tr>
</tbody>
</table>

NEWLY PROPOSED CLASSIFICATION STRUCTURE

The previously described classification tree was based on database descriptors and on bibliographic data where no full text was available. As a final and concluding step we attempted to create our own classification tree regardless of the topical subjects of the existing thesauri, and basing on full documents.

To this end we selected some 140 full-text documents (journal articles and books). We employed the period of 1997 onwards and extracted all the full-text journal articles that were
available via electronic publishers Kluwer, Springer, ScienceDirect, and DOAJ (Directory of Open Access Journals). We also used some web pages. In addition we analysed books that were available in the libraries of the Biotechnical Faculty.

With the help of a thorough analysis of these full original documents we created, based on our expert knowledge of the field of functional foods, the most appropriate classification tree regardless of the above already existing classification tree.

We analysed these original documents, and the data extracted and structured for 35 biologically active compounds. We then arranged these data into a tree system according to the authoritative chemical classification principles. This experimental system is presented in Figure 6. The comparison of Figure 5 (higher information density keywords - 4 or more occurrences) and Figure 6 (keywords extracted from original documents) reveals a substantial similarity between both classification tree structures. This is especially the case for branches of higher hierarchical order. However, some individual biologically active compounds from Figure 6 are not present in Figure 5 (e.g. alliin and sinirgin). This is the case with some more recently investigated active compounds which have not, as yet, been accepted as descriptors in the existing databases and the respective thesauri.

We can thus assume that in order to choose a workable classification tree, the existing database keywords (descriptors) with a higher information density can be successfully used as a basis. Such a classification system, however, still needs to be complemented by some additional keywords extracted from the original full-text documents (e.g. articles).
Figure 5: Classification of biologically active compounds with keyword frequency of four or more, based on existing terms (white areas), and newly supplemented terms (grey areas).
CONCLUSIONS

The proposed bibliometric methodology, based on keyword analysis and structuring of data into hierarchical tree-systems, can be used for assessment of bibliographic databases and
identification of research trends. An improved, upgraded, and better defined tree-structured classification scheme may enhance thesaurus-based database retrieval in newly emerging scientific fields.

The proposed methodology was tested on functional foods as an example of a newly emerging scientific field. FSTA database was selected for further detailed keyword analysis on account of this database having by far the highest number of relevant records. An analysis and structuring of the existing FSTA-related keywords into hierarchical tree-systems indicate that keyword indexing in this database sometimes differs from principles of the standard chemical classification. Especially, broader terms (broader-concept keywords, closer to the root node) are in certain cases used less frequently and less systematically.

The database and document analysis confirmed our previous assumption of Foods, Classification, Legislation, Diseases, Processes, Physiological Effects, and Properties as being the most clearly profiled topical groups within the broader field of functional foods. We created a classification tree based on the above groups and the descriptors derived from database thesauri. The thus obtained classification tree, however, lacked several relevant keywords so we upgraded and "improved" this existing tree with some additional keywords according to the appropriate chemical structure. We also designed a completely new "ideal" classification tree based on full-text documents and our knowledge of the field.

If we compare the existing tree with the model tree, we may, with some caution, conclude that the existing tree describes sufficiently well the field of functional foods providing we use the existing keywords with high information density (i.e. high frequency of occurrences). We would still strongly suggest to enhance the existing hierarchical tree with supplementary terms which can be harvested from authoritative resources. This can make up for inaccuracies and inconsistencies of the existing keyword indexes.

This methodology of database analysis may serve to improve database patterns, especially with regard to information retrieval. Greater emphasis thus needs to be given to information technologies and professional scientific information services which review, update and improve existing thesauri. Retrieval may be enhanced only by improved classification trees and descriptor structures, what is especially important in newly emerging scientific fields which usually lack systematic and consistent terminological description.

References


FSTA Thesaurus (2004), Compiled by Linda Merrywether; assisted by members of the IFIS Publishing editorial and production team. Reading: IFIS – International Food Information Service. p. 413.
Data structuring and classification in newly emerging scientific fields

(An information method based on bibliometric analysis of keywords used for data structuring and classification in newly emerging scientific fields. Example: functional food)

Simona JUVAN
Information specialist. Central Biotechnical Library, Biotechnical Faculty, University of Ljubljana, Slovenia.

Tomaz BARTOL
Assistant Professor of Information Science at the Department of Agronomy, Biotechnical Faculty, University of Ljubljana, Slovenia.

Bojana BOH
Professor of Scientific and Technological Informatics at the Department of Chemical Education and Informatics, Faculty of Natural Science and Engineering, University of Ljubljana, Slovenia.

Authors’ version of Research paper
Published in: ONLINE INFORMATION REVIEW, 2005, vol. 29, no. 5, p. 483-498

Keywords
bibliometric analysis, bibliographic databases, data processing, indexing, subject headings, thesaurus, thesauri, data structures, keyword analysis, structuring data into systems, classification, functional food, foods, human nutrition, bibliometrics, scientometrics, data mining

Abstract
The article addresses a methodological procedure based on keyword analysis and structuring of data into information systems in the field of functional foods as a newly emerging scientific field within the broader scope of food sciences and technology. Experiment was based on selection of a research field or research subject, selection of search profile, selection and processing of relevant databases, keyword analysis, and arranging of data (keywords) according to tree-structures. Keyword analysis was employed to identify narrower research fields within the broader scientific field. Structuring of data into systems was used to systematically classify the terms within the particular narrow field. Keywords with higher and lower frequency were identified. A classification tree was set up, based on keywords (thesaurus-based descriptors) extracted from FSTA database available online. The tree was supplemented and upgraded with some additional topical terms, that have not as yet been included in the existing thesaurus. To serve as a comparison a completely new classification tree was designed based on online full-text data. The comparison of the two trees suggests the previous existing tree as being sufficiently accurate to represent the field of functional foods, providing that it be upgraded with some additional terms. A more accurate classification should improve thesauri and consequently enhance retrieval in international databases.

INTRODUCTION

Rapid growth of information and availability of data from an increasing amount of new sources have boosted possibilities of data and information exploration, and, consequently, of identification of research trends and patterns in a particular area of knowledge. In the field of library and information science, quantitative studies of information have been marked by evolution of bibliometrics (Sanz-Casado et al., 2002). Bibliometric analyses can thus be very useful in following research trends (Mela et al., 1999).

In most bibliographic or full text databases, users can search either within free text, or with controlled terms (keywords or descriptors). Controlled terms allow a searcher an improved retrieval precision of documents on a selected topic, regardless of the words that the authors of a particular document had assigned to the topic. Indexing, however, generally depends on human analysis of the subject. Indexers read the material, locate the best terms in a thesaurus, and assign the terms that describe a document’s meaning. Common types of subject description in databases are usually arrayed from broad to specific.
Studies of indexing show significant variation in the use of terms chosen by different indexers to represent the same topic or document (Bertrand et al., 1995). Suraud et al. (1995) observed non-existence of well-defined keywords in newly emerging fields, what makes bibliographic searches difficult. Bates et al. (1993) call for development in the structures of thesauri and in design of online information systems. Hurt (1997) maintains that it is necessary to renew and expand indexing and classification systems. Soergel et al. (2004) also point out that existing classification schemes and thesauri lack in well defined semantics and structural consistency.

The above indicates that the existing systems do not necessarily allow for sufficiently accurate and correct information retrieval so some further development and improvement of search systems might be useful. Newly emerging scientific fields in particular need a more correct classification and better inclusion of pertaining terms in international bibliographic thesauri.

In our research we will try to characterise a newly emerging scientific field (functional foods) with regard to some principles of database indexing. We will identify and assess the existing indexing terms and existing classification tree in leading bibliographic databases in the field. We will complement and thus upgrade, where necessary, the existing indexing terms (descriptors) with some additional suitable terms which have hitherto not been assigned to the documents or not as yet been proposed as indexing terms in the existing thesauri. Also, we will analyse some selected online full-text documents in order to create a completely new classification tree. We will compare the existing, the upgraded, and the newly proposed classification trees.

The following methodological approach is offered as contribution to facilitate and improve the process of creating classification structures and thesauri in newly emerging scientific fields, and consequently to improve database quality and retrieval of documents.

MATERIALS AND METHODOLOGY

Our methodology was based on three concepts: (1) an information method to achieve value-added processing of bibliographic databases in science and technology (Kardos et al., 2000), complemented by (2) a bibliometric analysis (Mela et al., 1999) to define research trends and obtain keywords, and (3) a method to structure the data into systems (Kornhauser, 1989) for the classification of keywords.

The procedure was performed through the following steps (Figure 1): (1) selection of a research field, (2) definition of a search profile – search query, (3) selection and processing of an appropriate database, (4) extraction and analysis of keywords from the records retrieved with a well defined search query, and (5) arrangement of keywords into a tree-structured classification system.
Selection of a research field

As the first step a specific "research field of interest" was defined and the relevant scientific terms were identified.

Definition of a search profile

After the research field had been selected, an appropriate search profile or query was set up, and the time scope for the retrospective search and retrieval was defined. The search query typically consists of a combination of keywords and a time series, and some other possible searchable parameters.

In the case of a relatively new scientific field, the search query needs to include all potential synonyms, and needs to be based on all entries, not only on the descriptor entry, to minimise the risk of loosing documents on account of potential terminological inconsistencies in newly developing concepts.

This is in accordance with the observation of Suraud et al. (1995) whereby uncontrolled entries like title, abstract and free terms better reflect evolution of a concept and might be more easily adaptable to an emerging concept, as opposed to descriptors and classification code entries, which offer a relatively rigid access to a document.

Selection and processing of appropriate databases

In the third step, the most relevant database was selected. Within a group of potentially interesting databases, selection can be conducted on the basis of a search preview tool or
similar search options. In addition to the number of relevant records, other criteria can be considered (Kardos et al., 2000), such as database content, availability, and price.

**Extraction and analysis of keywords**

In the fourth step, keywords were extracted from the retrieved records, and were classified into groups according to main research trends within the scope of the selected research topic. This classification is best made on the basis of *title* and *abstract* of a record.

In our keyword analysis, a method by Mela et al. (1999) was used, where keywords were defined as comma-separated items of one or more words. Different keywords with identical meaning (synonyms) were identified "manually", and were further assigned a single-keyword-meaning. The keywords were then arranged into sub-groups of related terms.

**Structuring of keywords into a tree-structured classification system**

In the fifth step, arranging of keywords according to tree structures was carried out in the following order: (1) arranging of parameters into large sets of data - in our case an identification of parameters whereby keywords can be classified into sub-groups (Kornhauser, 1989); (2) defining hierarchy of parameters - arranging of keywords according to a broader or narrower (more specific) meaning, and (3) arranging of data into a tree-structured system, i.e. developing a logical classification system, which includes all extracted keywords.

The method allowed us to build a logical classification system according to the standard principle of arranging the terms from a broad concept to more specific, narrower concepts.

As a research material and as a terminological reference we used online bibliographic databases, Web pages, full-text journals by major scientific providers, and books in libraries.

**RESULTS AND DISCUSSION**

In the case of our research a method itself can also be interpreted as a part of the results of the analysis. We thus present some of the methodological issues also in the following field.

**Selection of a research field**

The concept of *functional foods* was selected as an example of a newly emerging scientific field. According to an accepted definition, functional foods are foods which are consumed as a part of normal daily diet, and which have health promoting benefits and reduce risk of diseases (Diplock et al., 2000; Roberfroid, 2000).

**Definition of a search profile – search query**

The concept of *functional foods* is not identical to the notions of *medical foods* or *dietary supplements*. It overlaps, however, with the notions of *foods for special dietary uses* and *fortified foods*. Functional foods may in some cases also be expressed as *nutraceuticals*, *health foods* or *novel foods* (Kwak et al., 2001).
The 2004 FSTA (Food Science and Technology Abstracts) Thesaurus positions functional foods in the following inter-term relationship: Synonym: *nutraceutical foods*, Related terms: *designer foods, health, medical foods, nutrition, phytochemical* and Broader term: *novel foods*.

Our search profile involved the 1990 – 2002 period, and entailed the following terms: *functional food* or *nutraceutical food* (employing right truncation). The retrieval with these terms was conducted in abstracts (AB), document titles (TI), and database keywords or descriptors (DE). Further examination, however, involved only the extraction and analysis of keywords from the Descriptor field.

**Selection and processing of databases**

The truncated search query *functional food* was employed to identify databases with the highest number of records in order to select the most appropriate database for further analysis.

The growth of publications on functional foods in 1990-2003 is presented in Figure 2. In most databases a rapid growth of documents can be noticed after 1997, what corresponds with the general recognition and acceptance of the definition of the term *functional food*. After 2001 the numbers of new publications show signs of some decrease. This may perhaps be explained by the fact that the research field *functional food* has now somehow stabilised what is mirrored also by the number of newly published documents.

![STN Free Search Preview: functional food* (1990-2003)](image)

Figure 2: STN Free Search Preview – growth of records in bibliographic databases that cover food science and technology research; search profile: *functional food* in the period 1990-2003; abbreviations are explained in Table 1 (processed 7.1.2005)
Table 1: Examples of bibliographic databases containing data relevant to the functional food research (selected from the STN International host – Food Science and Technology Cluster, processed 21.4.2004)

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Database Type</th>
<th>Coverage</th>
<th>File Size April 2004</th>
<th>Producer</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>CABA</td>
<td>Bibliographic database</td>
<td>1973 to the present</td>
<td>More than 4,500,000 records</td>
<td>CABI Publishing, UK</td>
<td>Agriculture and related science including veterinary medicine, human nutrition, developing countries, leisure, recreation, and tourism.</td>
</tr>
<tr>
<td>CAPLUS</td>
<td>Bibliographic database</td>
<td>1907 to the present</td>
<td>More than 23,590,000 records</td>
<td>Chemical Abstracts Service, USA</td>
<td>All areas of chemistry and chemical engineering. The major subdivisions are: Applied Chemistry and Chemical Engineering, Biochemistry, Macromolecular Chemistry, Organic Chemistry, Physical, Inorganic, and Analytical Chemistry.</td>
</tr>
<tr>
<td>FOMAD</td>
<td>Bibliographic database</td>
<td>1982 to the present</td>
<td>More than 181,000 records</td>
<td>FIZ Karlsruhe, Germany</td>
<td>Detailed analyses of international food and drink markets, identifying key market players, highlighting new product launches, assessing consumer attitudes and retail trends, and tracking company news.</td>
</tr>
<tr>
<td>FROSTI</td>
<td>Bibliographic database</td>
<td>1972 to the present</td>
<td>More than 549,000 records</td>
<td>The Leatherhead Food Research Association, UK</td>
<td>All aspects of the food and drink industry are covered.</td>
</tr>
<tr>
<td>FSTA</td>
<td>Bibliographic database</td>
<td>1969 to the present</td>
<td>More than 600,000 records</td>
<td>IFIS Publishing, UK</td>
<td>All areas which relate to food and beverages including the latest research, new product development, quality assurance and control, food safety, food composition, regulations, consumer trends.</td>
</tr>
<tr>
<td>JICST-EPlus File</td>
<td>Bibliographic database</td>
<td>1985 to the present</td>
<td>More than 4,723,000 records</td>
<td>Japan Science and Technology Agency (JST)</td>
<td>The literature published in Japan on all fields of science, technology, and medicine.</td>
</tr>
<tr>
<td>PROMT</td>
<td>Multiple-industry database</td>
<td>1978 to the present</td>
<td>More than 11,234,000 records</td>
<td>Gale Group, USA</td>
<td>International coverage of companies, products, markets, and applied technologies for approximately 60 manufacturing and service industries.</td>
</tr>
</tbody>
</table>

Explanation of database acronyms:
CABA: CAB Abstracts
CAPLUS: Chemical Abstracts Plus Database
FOMAD: Foodline®: Market Sight
FROSTI: Foodline®: Science Sight
FSTA: Food Science and Technology Abstracts
JICST-EPlus File: JICST file on Science, Technology, and Medicine in Japan
PROMT: Predicasts Overview of Markets and Technology
Other international bibliographic databases, which also include records on food science and technology research (e.g. AGRIS and MEDLINE) contained fewer than 1000 functional-foods-related records in the observed period, and were consequently excluded from further analysis.

FSTA database was selected for further analysis as the principal source of relevant records. The 1990-2002 period query "functional food* or nutraceutical food*" retrieved 1555 different records (1164 journal articles, 133 reviews, 79 conference proceedings, 77 patents, and 102 other type documents such as thesis, books, reports). A rapid increase of journal articles can be especially noticed in the more recent period (Figure 3).


EXTRACTION OF THE EXISTING KEYWORDS

Keyword analysis

Our bibliometric examination harvested 1555 initial records. The total corpus of keywords was 8367, including multiple occurrences of the same keyword in different records. The total of different original database-assigned keywords was 1223. Prior to the examination we proposed, according to our expertise in the field, the following possible topical groups: foods, classification (of biologically active compounds), properties (of biologically active compounds), physiological effects, foods (in general sense), diseases, legislation, and processes.
Group *Foods* represents those documents where a food (a fruit, vegetable, meat product etc.) was referred to in general sense. Group *Classification* contains those keywords which represent a specific biologically active compound, such as calcium, inulin, lactoferrin, folic acid. Group *Properties* represents some physico-chemical and organoleptic properties (e.g. physical properties, colour, solubility, texture) of foods. Our topical clustering of keywords and assessment of documents also yielded some other documents with the prevalence of documents described as bibliography which were usually some type of indexes, and which we excluded from topical subject analysis. The groups also presented in Table 2.

Table 2: Keyword structuring of functional foods into main topical groups with regard to number of different keywords, and database keyword occurrence.

<table>
<thead>
<tr>
<th>Group</th>
<th>Different terms (keywords)</th>
<th>Keyword occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOODS (e.g. yoghurt, eggs, chocolate)</td>
<td>419</td>
<td>3077</td>
</tr>
<tr>
<td>CLASSIFICATION (e.g. calcium, inulin, lactoferrin, folic acid)</td>
<td>325</td>
<td>1761</td>
</tr>
<tr>
<td>LEGISLATION (e.g. regulations, policy, standards)</td>
<td>124</td>
<td>1128</td>
</tr>
<tr>
<td>DISEASES (e.g. diabetes, osteoporosis, colon cancer)</td>
<td>79</td>
<td>869</td>
</tr>
<tr>
<td>PROCESSES (e.g. chromatography, encapsulation, freezing)</td>
<td>179</td>
<td>767</td>
</tr>
<tr>
<td>BIBLIOGRAPHY (e.g. patents, reports)</td>
<td>10</td>
<td>375</td>
</tr>
<tr>
<td>PHYSIOLOGICAL EFFECTS (e.g. immunology, digestibility, antitumour activity)</td>
<td>41</td>
<td>233</td>
</tr>
<tr>
<td>PROPERTIES (e.g. physical properties, colour, solubility, texture)</td>
<td>36</td>
<td>129</td>
</tr>
<tr>
<td>OTHER</td>
<td>10</td>
<td>28</td>
</tr>
</tbody>
</table>

Frequency or occurrence was calculated for each keyword, and was specified as a parameter of information density. The term information density is characterised as a frequency of occurrence of a separate search term in a separate search field (Boh, 1996; Kardos and Boh, 2000), and can be expressed as a percentage (e.g. 20%, if a search term appears in 20% of records), or as a number.

**Structuring of keywords into a classification tree system**

Our further analysis involved only the keywords from the above group that was tagged as *Classification*. Namely, this group comprised only the biologically active compounds, e.g. calcium, beta carotene, inulin, lactoferrin etc. Because of the existing inconsistent chemical classification of biologically active compounds in functional foods the method of structuring of data into systems (Kornhauser, 1989) was used to propose an improved and more consistent classification tree.

In most cases we used the FSTA Thesaurus (2004) as a term of reference. In some cases, however, this thesaurus was not suitable because it is not always in accordance with the standard chemical classification of active compounds.

An example of insufficiency of this thesaurus is the arrangement of carbohydrates which are generally a well established chemical group. Cummings et al. (1997) divide carbohydrates into monosaccharides (glucose and fructose), disaccharides (sucrose, lactose, etc.), sugar alcohols (sorbitol, maltitol, etc.), oligosaccharides (maltooligosaccharides and nondigestible oligosaccharides) and polysaccharides (starch and non starch polysaccharides). FSTA Thesaurus (2004), on the other hand, divides carbohydrates into *bifidus factors, glycosides,*
oligouronides, polyols, polysaccharides and sugars (amino sugars, disaccharides, oligosaccharides, reducing sugars, etc.). In this case we gave priority to the afore-mentioned classification by Cummings to obtain a more accurate chemical classification and to avoid having too many hierarchical levels.

UPGRADING OF THE EXISTING KEYWORD STRUCTURES

During the process of data structuring, some keywords were excluded, such as keywords that did not represent specific compounds (e.g. keywords ingredients, composition) or some general group that could not be placed into a specific chemical group. Those keywords that generally stood for a functional group we arranged into a corresponding chemical group.

We identified synonyms (e.g. vitamin C / ascorbic acid) and then merged these synonyms as a single term. We structured the remaining keywords from the group Classification into subgroups according to the chemical nature of a particular compound. We thus created additional two main "chemical" branches of the tree system Inorganic compounds and Organic compounds. We added these two groups to the already existing field of Microorganisms (Figure 4). Figure 4 represents the left fraction (main branches) of the Figure 5. White fields represent the existing thesauri-derived keywords with high information density. Grey fields in the figures represent additional categories which we created in order to better describe and explain chemical structure of the field of functional foods.

![Classification Tree](image)

Figure 4: Main branches of classification tree of biologically active compounds in functional foods, based on existing (white areas) and supplemented (grey areas) terms.

In some cases we thus created a new category (according to chemical classification) that had not existed in the original classification structure or the initial set of keywords and which had
not been obtained during the bibliometric analysis. We defined such a group with the help of thesaurus or some other source of classification of chemical compounds. This consequently resulted in a more consistent classification in our experimental database.

While counting keywords we observed that those with higher frequency (higher information density) are usually represented as main branches on the left side of the classification tree (e.g. minerals, vitamins, antioxidants), while keywords with a lower frequency, such as individual biologically active compounds (e.g. calcium, vitamin A, quercetin, zeaxantin), represent the branches of lower hierarchical order on the right side (Figure 5).

We noticed, however, a few exceptions. In some cases high-frequency keywords come about in lower branches, such as the keyword inulin with 24 occurrences, and which had been assigned the position of the sixth branching (the furthest from the root node in the Classification). Most other keywords with similar occurrences were placed in the second or third branching of the tree. Another example of an exception is the keyword fructans with four occurrences. It was positioned in a branch closer to the root node than inulin even though fructans are a broader concept than inulin (according to the FSTA Thesaurus) and should occur more frequently.

The analysis and structuring of keywords indicate that the indexing of the FSTA database frequently differs from the standard chemical classification. Some records which contain keywords glucose and fructose represent an example of inconsistence. Both substances can be chemically classified as monosaccharides. However, this term (monosaccharides) was not assigned as a descriptor to the respective records in the FSTA database, even though the term monosaccharides does feature as a descriptor in the FSTA thesaurus. In some cases, but again not consistently, there occurred the keyword sugars. From the chemical point of view, where carbohydrates are usually divided into monosaccharides, disaccharides, oligosaccharides and polysaccharides, for the keywords glucose and fructose the group term monosaccharides would be more appropriate than the term sugars (as sugars can be monosaccharides or disaccharides) or at least it would be better to use both keywords: sugars and monosaccharides. This leads to conclusion that indexing of this database differs from some principles of the standard chemical classification and that the broader terms are in certain cases used less systematically.

Theses findings concur with some previous research. For instance, Cohen et al. (1979) noticed that relations between broader and narrow terms in the FSTA thesaurus were not always accurate. Weintraub (1992) compared searches in agricultural databases such as AGRICOLA, AGRIS and CAB with regard to the topic of alternative agriculture and established that standardised terminology was lacking. Hutchinson et al. (1996) pointed out lack of standardisation in spelling and of certain terminology as the main weakness of FSTA thesaurus. Also, Bartol et al. (2000) stressed that in some databases, such as AGRIS, CAB, FSTA and MEDLINE, many different terms exist for the same topic, and these can sometimes be used as a narrow and sometimes as broader term without inter-database consistency.

One of the methodological problems in structuring of keywords into a classification tree is to determine which keywords of higher (broader) information density can substitute some other more specific (narrower) keywords. We tested the above by extracting and structuring only the keywords with the frequency of four or more occurrences from the group of biologically active compounds. The experiment indicated that the significant groups of biologically active compounds (e.g. vitamins, minerals, antioxidants) and important individual compounds (e.g.
calcium, vitamin E, lactoferrin) are still present in the tree system as shown in Figure 4. This result suggests that it might be sufficient to use only the keywords with higher information density, however, with some restriction. The usage of only higher-information-density keywords will capture most broader chemical categories. On the other hand, some documents may be lost, especially those describing novel compounds and not as yet clearly defined new scientific terms.

An example are the organic sulphur compounds found in garlic. Garlic is a functional food with more than 30 organic sulphur-containing compounds, such as alliin as a biologically active compound (Markowitz et al., 2003). Our study found out that there was only one keyword representing the organic sulphur compounds (isothiocianates with 2 occurrences). Alliin was not present among the keywords, even though alliin does come about as a descriptor in the FSTA Thesaurus, as showed another search in FSTA database (Table 3) with different search profiles containing keywords garlic, alliin and functional food*. Records that contained the term alliin had no correlation to the term functional food. An analysis based strictly on keywords might conclude that organic sulphur compounds have no significance as biologically active compounds. Such a conclusion would be incorrect because it would be based on deficient indexing of documents due to inadequate knowledge on the part of the indexing staff. The results therefore indicate that in certain cases indexing in FSTA database is not conducted in comprehensive agreement with the principles of chemical classification. Such deficiencies are not a characteristic of FSTA thesaurus only. Most thesauri exhibit some level of shortage or subjectivity with regard to terminological representation. Moreover, this fact applies not only to new emerging scientific research fields but also to basic fields, e.g. chemistry in general.

Table 3: Documents in FSTA database (1990-2002) containing different combinations of keywords garlic, alliin, and functional food*.

<table>
<thead>
<tr>
<th>Search profile</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>alliin</td>
<td>63</td>
</tr>
<tr>
<td>alliin in de</td>
<td>42</td>
</tr>
<tr>
<td>garlic</td>
<td>1013</td>
</tr>
<tr>
<td>garlic in de</td>
<td>483</td>
</tr>
<tr>
<td>alliin and garlic</td>
<td>47</td>
</tr>
<tr>
<td>(alliin and garlic) in de</td>
<td>19</td>
</tr>
<tr>
<td>alliin and functional food*</td>
<td>0</td>
</tr>
<tr>
<td>garlic and functional food*</td>
<td>16</td>
</tr>
<tr>
<td>(alliin in de) and functional food*</td>
<td>0</td>
</tr>
<tr>
<td>(garlic in de) and functional food*</td>
<td>1</td>
</tr>
</tbody>
</table>

NEWLY PROPOSED CLASSIFICATION STRUCTURE

The previously described classification tree was based on database descriptors and on bibliographic data where no full text was available. As a final and concluding step we attempted to create our own classification tree regardless of the topical subjects of the existing thesauri, and basing on full documents.

To this end we selected some 140 full-text documents (journal articles and books). We employed the period of 1997 onwards and extracted all the full-text journal articles that were
available via electronic publishers Kluwer, Springer, ScienceDirect, and DOAJ (Directory of Open Access Journals). We also used some web pages. In addition we analysed books that were available in the libraries of the Biotechnical Faculty.

With the help of a thorough analysis of these full original documents we created, based on our expert knowledge of the field of functional foods, the most appropriate classification tree regardless of the above already existing classification tree.

We analysed these original documents, and the data extracted and structured for 35 biologically active compounds. We then arranged these data into a tree system according to the authoritative chemical classification principles. This experimental system is presented in Figure 6. The comparison of Figure 5 (higher information density keywords - 4 or more occurrences) and Figure 6 (keywords extracted from original documents) reveals a substantial similarity between both classification tree structures. This is especially the case for branches of higher hierarchical order. However, some individual biologically active compounds from Figure 6 are not present in Figure 5 (e.g. *alliin* and *sinirgin*). This is the case with some more recently investigated active compounds which have not, as yet, been accepted as descriptors in the existing databases and the respective thesauri.

We can thus assume that in order to choose a workable classification tree, the existing database keywords (descriptors) with a higher information density can be successfully used as a basis. Such a classification system, however, still needs to be complemented by some additional keywords extracted from the original full-text documents (e.g. articles).
Figure 5: Classification of biologically active compounds with keyword frequency of four or more, based on existing terms (white areas), and newly supplemented terms (grey areas).
CONCLUSIONS

The proposed bibliometric methodology, based on keyword analysis and structuring of data into hierarchical tree-systems, can be used for assessment of bibliographic databases and
identification of research trends. An improved, upgraded, and better defined tree-structured classification scheme may enhance thesaurus-based database retrieval in newly emerging scientific fields.

The proposed methodology was tested on functional foods as an example of a newly emerging scientific field. FSTA database was selected for further detailed keyword analysis on account of this database having by far the highest number of relevant records. An analysis and structuring of the existing FSTA-related keywords into hierarchical tree-systems indicate that keyword indexing in this database sometimes differs from principles of the standard chemical classification. Especially, broader terms (broader-concept keywords, closer to the root node) are in certain cases used less frequently and less systematically.

The database and document analysis confirmed our previous assumption of Foods, Classification, Legislation, Diseases, Processes, Physiological Effects, and Properties as being the most clearly profiled topical groups within the broader field of functional foods. We created a classification tree based on the above groups and the descriptors derived from database thesauri. The thus obtained classification tree, however, lacked several relevant keywords so we upgraded and "improved" this existing tree with some additional keywords according to the appropriate chemical structure. We also designed a completely new "ideal" classification tree based on full-text documents and our knowledge of the field.

If we compare the existing tree with the model tree, we may, with some caution, conclude that the existing tree describes sufficiently well the field of functional foods providing we use the existing keywords with high information density (i.e. high frequency of occurrences). We would still strongly suggest to enhance the existing hierarchical tree with supplementary terms which can be harvested from authoritative resources. This can make up for inaccuracies and inconsistencies of the existing keyword indexes.

This methodology of database analysis may serve to improve database patterns, especially with regard to information retrieval. Greater emphasis thus needs to be given to information technologies and professional scientific information services which review, update and improve existing thesauri. Retrieval may be enhanced only by improved classification trees and descriptor structures, what is especially important in newly emerging scientific fields which usually lack systematic and consistent terminological description.

References
