Journal of the American Society for Information Science and Technology

The presence of the keywords given by authors of scientific articles in databases descriptors

Journal:	*Journal of the American Society for Information Science and Technology*
Manuscript ID:	draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Keywords:	indexing < organization of information < information operations < (activities and operations), keywords < index terms < terms < terminology < information representations < (knowledge and information), descriptors < index terms < terms < terminology < information representations < (knowledge and information)

The presence of the keywords given by authors of scientific articles in databases descriptors

Isidoro Gil-Leiva
University of Murcia, Faculty of Communication and Documentation, Spain, isgil@um.es

Adolfo Alonso Arroyo
University of Valencia, Faculty of Medicine, Spain, adolfo.alonso@uv.es

ABSTRACT

This paper analyses the keywords given by authors of scientific articles and the descriptors assigned to the articles in order to ascertain the presence of the keywords in the descriptors. 640 INSPEC, CAB abstracts, ISTA and LISA database records were consulted. After detailed comparisons it was found that keywords provided by authors have an important presence in the database descriptors studied, since nearly 25% of all the keywords appeared in exactly the same form as descriptors, with another 21% while normalized, are still detected in the descriptors. This means that almost 46% of keywords appear in the descriptors, either as such or after normalization. Elsewhere, three distinct indexing policies appear, one represented by INSPEC and LISA (indexers seem to have freedom to assign the descriptors they deem necessary); another is represented by CAB (no record has fewer than four descriptors and, in general, a large number of descriptors is employed; in contrast, in ISTA, a certain institutional code towards economy in indexing, since 84% of records contain only four descriptors.

Keywords: Keywords ; Descriptors ; Indexing ; Automatic indexing ; Journal papers ; Databases ; Comparative study ; INSPEC ; CAB ; ISTA ; LISA.

1. Introduction

Indexing is the procedure applied to the content of documents and the questions to select those concepts which best represent them, and thus facilitate storing and retrieval. ISO norm 5963:1985 recommends that during analysis of text documents "special attention be paid" to titles, abstracts, summaries or content tables, introductions, opening paragraphs of chapters or sections, conclusions, illustrations, diagrams, tables and captions, and underlined or highlighted words or sentences.

Keyword is "a word or group of words, possibly in lexicographically standardized form, taken out of a title or of the text of a document characterizing its content and enabling its retrieval." (ISO 5963-1985).

While not seeking to be exhaustive, we can point out that research into keywords has dealt with a variety of subject matters:

▪ Retrieval efficiency:

- Gross and Taylor (2005), on the debate on whether it is necessary to assign subject headings in library catalogues or to use keywords for retrieval, studied what effect keywords have on retrieval if catalogues do not include the field subject heading.

- Taghva et al. (2004) explored the use of manually assigned keywords for query expansion with interactive tools.;
- Voorbij (1998) analysed the value of subject matter descriptors and keywords from titles in subject matter searches; and
- Tillotson (1995) investigated the possibilities of OPAC interfaces for search by keywords and controlled vocabulary. They also performed several experiments on the relevance of searching by keywords.

▪ Use by authors and editors:

- Hartley and Kostoff (2003) reviewed journals from various disciplines to verify which habitually provided keywords. They also asked 35 editors of scientific journals to explain the advantages and drawbacks of using keywords.

- Gbur and Trumbo (1995) put forward ten recommendations for choosing suitable keywords for journal papers, along with suggestions for preparing informative titles and useful abstracts both for readers and database producers.

▪ Meta tag keywords:

- Craven (2005 and 2004) studied meta tag keywords of websites for the 19 languages most commonly present on the web, and determined the effect of website edition tools on meta tag keywords.
- Alimohammadi (2004) calculated the presence of meta tag keywords in 346 websites of Iran.

▪ Automatic extraction:

- The use of different methodologies and algorithms to obtain keywords has been the subject of repeated research in recent decades (Lancheng, 2005; Jones and Paynter, 2002; Boger et al., 2001 or Turney, 2000, to name but some recent studies).

▪ Comparison of keywords in titles, abstracts and texts with assigned descriptors.

- Ansari (2005) compared the descriptors assigned to 506 doctoral theses from the Department of Indexing of the Iran University Central LIbrary using the keywords from the titles of these theses.

- Gil-Leiva and Rodríguez Muñoz (1997) compared keywords in titles and abstracts from 450 scientific articles from Sciences, Social Sciences and Medical Sciences using the descriptors assigned in three Spanish databases maintained by professional indexers.

With the exception of our own paper (Gil-Leiva and Alonso-Arroyo, 2005) we do not know of any other research which deals with the function that keywords provided by authors of scientific articles may or may not perform in professional indexing. The interest in verifying this possible influence is twofold. On the one hand, we will gain more knowledge of the intellectual process realized by the indexers and this understanding will serve for a possible integration into methodologies applied in

automatic indexing, which uses rules taken from human indexers. Authors' keywords could be used as well as titles, abstracts and texts for automatic indexing of articles.

In Gil-Leiva and Alonso-Arroyo (2005) we randomly selected 108 scientific journals which were proportionally distributed among Social Sciences and Humanities (36), Science and Technology (36) and Medical and Health Sciences (36). 10 articles, from different various years, were randomly selected from the 108 publications. Our final working sample was 1080 articles which fulfilled two conditions: they possessed keywords and they were included in the ISOC, ICYT and IME[1] databases. We subsequently contrasted the keywords provided by the authors of the articles with the assigned subject matter descriptors.

As we will see later, the results of this study show that the keywords given by the authors of scientific authors are directly or indirectly present in the subject matter descriptors assigned by professional indexers. Nevertheless, we considered that it was necessary to carry out further experiments on international databases in order to confirm the results obtained.

Thus, the aim of this paper is to calculate the direct (exact) presence or the indirect presence (after a minor normalization process) of the keywords given by the authors of scientific articles in the descriptors assigned by professional indexers. For this purpose, we chose the international databases INSPEC, CAB ABSTRACT, ISTA and LISA.

2. Materials and methods

This study was carried out using 640 scientific articles which fulfilled two conditions: they possess keywords given by the authors and are indexed in the databases mentioned. The articles belong to disciplines included in four databases: INSPEC (Physics, Electrical and Electronic Engineering, Computer Sciences, etc.), CAB Abstract (Agriculture, Forestry, Veterinary Science, Nutritional Sciences, etc.), ISTA (Information Science and related disciplines), and LISA (Libraries and Information Science). Annex 1 gives the journals and the years used in the study, while Annex 2 gives the document numbers of the 640 articles used in the study.

A table was drawn up for each of the 32 journals selected with the descriptors assigned to the 640 articles in the INSPEC, CAB, ISTA and LISA databases and the keywords given by the authors. The tables took the following structure:

> Source : minimum data for article identification, i.e. year, volume, number and first and last pages;

[1] These databases compile articles from the journals edited in Spain and they belong to the main research organism in Spain. El Consejo Superior de Investigaciones Científicas produces and distributes the ICYT (Science and Technology) database, which covers the period since 1979 to the present day. It indexes some 800 journals and incorporates more than 8.000 entries per year; lthe ISOC (Social Sciences and Humanities) database covering the period since 1975 performs the emptying of more than 2,000 journals and includes some 23,000 new references each year. In third place, the IME (Biomedicine) database covers the period since 1971. It indexes 321 journals and has some 200,000 entries. Finally, it should be observed that the three databases are maintained by professional indexers, specialized in each of the disciplines.

Keywords: list of the keywords provided given by the author;

Descriptors : list of the descriptors proposed by the indexers of each database;

Num. of keywords given by the author;

Num. of PCs Used: num. of keywords participating in the comparisons to find the exact coincidences and the normalizations between keywords and descriptors;

Num. of descriptors: num. of descriptors proposed by the indexers of each database;

Coincidences: num. of keywords coinciding exactly with the descriptors;

Normalized: num. of keywords which evoke concepts which also appear as descriptors, and which have apparently undergone only one normalization process;

Total : sum of the number of coincidences and the number of normalizations.

Tables I, II, III and IV show the procedure for gathering and comparing the keywords and descriptors. The author of the article in the first row of Table 1 gave four keywords (Global register allocation; Graph coloring; Linear scan; Binpacking) and the article was assigned 3 descriptors in the INSPEC database (Graph-colouring; Optimising-compilers; Storage-allocation). It can be observed that the keyword 'Graph coloring' also appears as a descriptor, and hence, in the column 'Coincidence' there is a 1; since no normalization process appears for any of the keywords in the Descriptors, a 0 appears in the column 'Normalization'. Hence, in the last column – the sum of 'Coincidence' and 'Normalization' there is a 1.

The column 'Normalization' quantifies to what extent one or several keywords proposed by an author evoke a concept later represented by one or more descriptors. We use the word 'evoke' in the sense of reminding or bringing to mind. This may be total or partial, i.e. a keyword by an author may bring to mind a complete concept or just a part, or in other words, a complete descriptor or a part of one. A value of equal to 1 was assigned for a seemingly complete reminder between keyword and descriptor, and 0.5 was assigned when it was partial. Table V shows various examples of this and Table VI gives the data for the journal Library Acquisitions Practice & Theory and the LISA database.

3. Results and discussion

Before presenting the results, it should be explained that the 24 journals studied here were reviewed in order to read the recommendations on keywords in the 'Instructions to authors'. In general, three to six keywords which cover the main issues dealt with in the paper are recommended. One of the journals includes the indication "which should complement the title but not repeat words in it". Annex 3 shows the most important of these.

3.1 Quantitative relation between the number of keywords given by authors and by descriptors

According to the data obtained, authors usually respect the guidelines in the 'Instructions to authors', as is confirmed in summarized form in Table VII and in greater detail in Annex 4. As mentioned, three to six words is the recommendation, although some authors include up to 20.

With regard to the descriptors assigned in the various databases, significant variations do appear for several aspects. On the one hand, while the total number of descriptors assigned is relatively similar in INSPEC (775), ISTA (646) and LISA (780), in contrast 1955 descriptors are assigned in CAB, much more than twice the number assigned in the other databases. The number of descriptors used per article differs likewise, which could be due to different indexing policies. Although the number of entries examined is not high, three apparent "models" of indexing are discerned. A first model in INSPEC and LISA, which leads to the indexing of 90% of the articles with between 2 and 9 descriptors; a second model, represented by CAB, where no article has fewer than four descriptors assigned to it, there is a compact band which has between 6 and 14, and then a substantial number from 15 up to 35 descriptors (there are 2 articles with 31 and 35, respectively.). Finally, the third model belongs to ISTA, where there seems to be a certain "economy" in the indexing since of the 160 entries, 135 have only four descriptors assigned to them.

In conclusion, these indexing policies mean that INSPEC, ISTA and LISA have an average of 4-5 descriptors per article, while in CAB the average stands at 12 descriptors. See Annex 4 for clear details.

3.2 Semantic relation between the number of keywords given by authors and the descriptors

Tables VIII, IX, X and XI provide examples which verify the lesser or greater presence of keywords in the descriptors. Table XII shows the total data for the four databases.

In Gil-Leiva and Alonso Arroyo (2005), the data obtained for the three databases studied were as follows: in IME 64.96% of the keywords were present in the descriptors; in ISOC, the figure was 60.48%, and 5n ICYT 5t was 58.18%. In the present study, the results were CAB (60.8%), LISA (42.2%), INSPEC (41.3%) and ISTA (37.89%). Despite the lower percentages, our hypothesis that keywords provided by the authors are an important source for indexing articles is confirmed.

It is therefore of use to take into account the keywords of the authors both when teaching indexing and in efforts to automate this process. The algorithms used in automatic indexing analyse a structured text partially or completely in order to propose a list of terms which represent the content of that text. These algorithms sometimes aim to simulate cognitive processes performed by human indexers, e.g. by giving more or less value to a word according to its position. This is the recommendation of ISO norm 5963/1985 devoted to '*Methods for examining documents, determining their subjects, and selecting indexing terms*'. Simulating intellectual procedures, automatic indexing systems are traditionally based on three sources to identify and value words or sentences, i.e. the titles of papers, the abstracts and the complete texts.

To the best of our knowledge, a review of the literature on automatic indexing does not reveal cases where keywords from the authors are used as a source. Titles have been dealt with by Kishida (2001); abstracts by Hmeidi, Kanaan and Evans (1997) and Ripplinger and Schmidt (2001); and titles and abstracts by Hersh and Hickam (1992) and Silvester, Genuardi and Klingbiel (1994). Complete texts have been studied by Gil-Leiva (1999 and 2003), Montejo Ráez (2002) or Ko, Park and Seo (2004).

SISA is an automatic indexing system (Gil-Leiva, 1999 and 2003) which handles titles, abstracts and complete text to propose indexing terms for the documents analysed. From the results obtained here, it is our intention to carry out the necessary changes for SISA to be able to take into account keywords by the authors as well. We will thus ascertain if improvements in results arise from the inclusion of this source.

4. Conclusions

Several aspects have come to light in this study. In the first place, there is a vacuum in the literature regarding the role of that keywords provided by authors of scientific articles do or do not play in the subsequent indexing of the texts. It has been seen that studies on keywords have dealt mainly with the efficiency of information retrieval, its use by authors and editors, the use of meta tag keywords or automatic extraction from texts.

Secondly, while the number of entries studied is not large, three indexing policies have been detected: one in which the indexer appears to be free to assign descriptors which he or she deems appropriate (INSPEC and LISA); a second one, represented by CAB, in which, in general, a large number of descriptors is employed. (in some cases up to 35); and, lastly, a certain type of institutional "economy" in indexing, with 84% of the entries analysed having only four descriptors.

Finally, the keywords given by the authors have an important presence in the database descriptors. 25% of all keywords handled in this study appear in exactly the same form as descriptors, while another 21% although they have undergone a normalization process, are still detected in the descriptors. This leads to around 46% of the keywords in the four databases appearing in the same or a normalized form as descriptors. These data confirm results by Gil-Leiva and Alonso Arroyo (2005), which means that keywords provided by authors are a valuable source of information for both human indexing and for automatic indexing systems of journal articles.

5. References

ALIMOHAMMADI, D. (2003). Meta-tag: a means to control the process of Web indexing. Online Information Review, 27 (4), 238-242.

ANSARI, M. (2001). Descriptors and title keywords: matching in medical PhD dissertations. Quarterly Journal of the National Library of the Islamic Republic of Iran, 12 (2), 23-33.

BOGER, Z. & KUFLIK, T. & SHOVAL, P. & SHAPIRA, B. (2001). Automatic keyword identification by artificial neural networks compared to manual identification by users of filtering systems. Information Processing and Management, 37 (2), 187-198.

CRAVEN, T. (2004). Variations in use of meta tag keywords by web pages in different languages. Journal of Information Science, 30 (3), 268-279.

CRAVEN, T. (2005). Web authoring tools and meta tagging of page descriptions and keywords. Online Information Review, 29 (2), 129-138.

GBUR, E.E. & TRUMBO, B.E. (1995). Key words and phrases—The key to scholarly visibility and efficiency in an information explosion. The American Statistician, 49, 29-33.

GIL-LEIVA, I. & RODRÍGUEZ MUÑOZ, J.V. (1997). Análisis de los descriptores de diferentes áreas del conocimiento indizadas en bases de datos del CSIC. Aplicación a la indización automática. Revista Española de Documentación Científica, 20 (2), 150-161.

GIL-LEIVA, I. (1999). La automatización de la indización de documentos. Gijón, Trea.

GIL-LEIVA, I. (2003). Sistema para la Indización Semi-Automática (SISA) de Artículos de Revista de Biblioteconomía y Documentación**.** II Jornadas de Tratamiento y Recuperación de Información, septiembre 2003, Leganés (Madrid), p. 228-232.

GIL-LEIVA, I. & ALONSO ARROYO, A. (2005). La relación entre las palabras clave aportadas por los autores de artículos de revista y su indización en las Bases de datos ISOC, IME e ICYT. Revista Española de Documentación Científica, 28 (1), 62-79.

GROSS, T. & TAYLOR, A.G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. College & Research Libraries, 66 (3), 212-230.

HARTLEY, J. & KOSTOFF, R.N. (2003). How useful are 'key words' in scientific journals?. Journal of Information Science, 29 (5), 433-438.

HERSH, W.R. & HICKAM, D.H. (1992). A comparison of retrieval effectivenes for three methods of indexing medical literature. The American Journal of the Medical Sciences, 303, 293-300.

HMEIDI, I., KANAAN, G. & EVENS, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabia documents. Journal of the American Society for Information Science, 48 (10), 867-881.

ISO 5963:1985 Documentation. Methods for examining documents, determining their subjects, and selecting indexing terms.

JONES, S. & PAYNTER, G.W. (2002). Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. Journal of the American Society for Information Science and Technology, 53 (8), 653-677.

KISHIDA, K. (2001). Statistical methods for automatically assigning classification numbers and descriptors based on title words of journal articles. Journal of Japan Society of Library and Information Science, 47 (2), 49-66.

KO, Y., PARK, J. & SEO, J. (2004). Improving text categorization using the importante of sentences. Information Processing & Management, 40 (1), 65-79.

LANCHENG, W. (2005). Theme information extraction of XMARC based on extended maximum matching algorithm. Journal of the China Society for Scientific and technical Information, 24 (1), 82-86.

MONTEJO RÁEZ, A. (2002). Towards conceptual indexing using automatic assignment of descriptors. Workshop in Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives. Málaga, Spain, May.

RIPPLINGER, B. & SCHMIDT, P. (2001). AUTINDEX: An Automatic Multilingual Indexing System. SIGIR, p. 452-452.

SILVESTER, J.P., GENUARDI, M.T. & KLINGBIEL, P.H. (1994). Machine-aided indexing at NASA. Information Processing & Management, 30 (5), 631-645.

TAGHVA, K., BORSACK, J., NARTKER, T. & CONDIT, A. (2004). The role of manually-assigned keywords in query expansion. Information Processing & Management, 40, 441-458.

TILLOTSON, J. (1995). Is keyword searching the answer? College and Research Libraries, 56 (3), 199-206.

TURNEY, P.D. (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2 (4), 303-336.

VOORBIJ, H.J. (1998). Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences. Journal of Documentation, 54 (4), 466-476.

Table I
INSPEC database

Sources	Articles	Keywords (KW)	Descriptors (DE)	Nº Kw	Nº Kw used	Nª DE	Exact	Normali	Total C+N
Sigplan Notices	Traub, 1998	1. Global register allocation 2. Graph coloring 3. Linear scan 4. Binpacking	1. Graph-colouring 2. Optimising-compilers 3. Storage-allocation	4	1	3	1	0	1
Performance Evaluation	Sheng, 2003	1. Admission control 2. Negotiation 3. Multimedia systems 4. Stochastic Petri-net 5. Queueing theroy	1. Multimedia-systems 2. Performance-evaluation 3. Petri-nets 4. Quality-of-service 5. Queueing-theory 6. Stochastic-processes	5	3	6	2	1	3

Table II
CAB ABSTRACT database

Sources	Articles	Keywords (KW)	Descriptors (DE)	Nº Kw	Nº Kw used	Nª DE	Exact	Normali	Total C+N
Water, Air & Soil Pollution	Papassiopi, 1999	1. EDTA 2. heavy metals 3. leaching 4. lead 5. soil remediation	1. removal 2. heavy-metals 3. calcareous-soils 4. polluted-soils 5. EDTA 6. leaching 7. soil 8. zinc 9. lead 10. pollution	5	4	10	4	0	4
Journal of Agricultural And Food Chemistry	Hornero, 2001	1. spectrophometry 2. capsicum annuum 3. carotenoids 4. paprika 5. oleoresin 6. quality	1. carotenoids 2. chemical-composition 3. determination 4. methodology 5. oleoresins 6. paprika 7. spectrophotometry 8. Capsicum 9. Capsicum-annuum	6	5	9	5	0	5

Table III
ISTA database

Sources	Articles	Keywords (KW)	Descriptors (DE)	Nº Kw	Nº Kw used	Nª DE	Exact	Normali	Total C+N
Information Processing & Management	Wildemuth, 2000	1. Factual databases 2. Medical students 3. Problem solving 4. Interface design	1. Computer-Interfaces 2. Databases 3. Information-Retrieval 4. Medical-Students 5. Problem-Solving 6. Clinical-Experience 7. Evaluation	4	4	7	3	0,5	3,5
Online Information Review	Fong, 2002	1. Internet 2. Research 3. Electronic publishing 4. Content analysis	1. Scholarly-communication 2. Document-access 3. Extracting 4. Practical-methods	4	0	4	0	0	0

Table IV
LISA database

Sources	Articles	Keywords (KW)	Descriptors (DE)	Nº Kw	Nº Kw used	Nª DE	Exact	Normali	Total C+N
Library Acquisitions : Practice & Theory	Shirk, 1994	1. Outsourcing 2. Processing 3. Contracts 4. Technical services	1. Technical-services 2. Contracting-out 3. Booksellers	4	2	3	1	1	2
International Journal of Information Management	Loebbecke, 1999	1. Information services 2. Electronic publishing 3. Electronic commerce 4. Multi-media	1. Electronic-publishing 2. Evaluation 3. Rentrop-Publishing	4	1	3	1	0	1

Table V
Examples of the use of 1 and 0.5 in Normalization column

Keywords from authors	Descriptors	Normalization
Internet	Internet	1
Libraries	Libraries	1
Parliament	Parliaments	1
Thesaurus construction	Thesauri Construction	1
Anglo-American Cataloguing Rules	AACR	1
Cathode ray tube (CRT) display	Cathode ray tube displays	1
Linear programming	Linear programming	1
Organic matter	Organic matter	1
Tomato	Tomatoes	1
Protein	Protein-content	0.5
Embryology	Embryos	0.5
Soil	Soil-pollution	0.5
Measurement	Statistics	0.5
Libraries	Digital-libraries	0.5
Cataloguing	Online-cataloguing	0.5
Departmental libraries	Academia-libraries	0.5
Democratization	Democracy	0.5
Red wine	Wines	0.5

Table VI
Data obtained for the journal *Library Acquisitions Practice & Theory*

Article	Num Kw	Num Kw Used	Num. DE*	Exact	Normal	Total E+N	% Exact	% Normal	% Total
1	4	1	2	0	1	1	0,00%	25,00%	25,00%
2	5	3	4	1	1,5	2,5	20,00%	30,00%	50,00%
3	4	3	4	1	2	3	25,00%	50,00%	75,00%
4	4	2	3	1	1	2	25,00%	25,00%	50,00%
5	3	3	8	1	2	3	33,33%	66,67%	100,00%
6	2	1	3	0	1	1	0,00%	50,00%	50,00%
7	6	3	5	2	1	3	33,33%	16,67%	50,00%
8	5	2	6	2	0	2	40,00%	0,00%	40,00%
9	4	2	4	2	1	3	50,00%	25,00%	75,00%
10	3	2	3	0	1,5	1,5	0,00%	50,00%	50,00%
11	7	2	3	2	0,5	2,5	28,57%	7,14%	35,71%
12	4	2	4	1	0,5	1,5	25,00%	12,50%	37,50%
13	6	3	5	1	2	3	16,67%	33,33%	50,00%
14	4	3	7	1	1,5	2,5	25,00%	37,50%	62,50%
15	5	3	6	0	2	2	0,00%	40,00%	40,00%
16	4	2	8	0	2	2	0,00%	50,00%	50,00%
17	3	1	4	1	0	1	33,33%	0,00%	33,33%
18	7	2	2	2	0	2	28,57%	0,00%	28,57%
19	4	1	4	2	0	2	50,00%	0,00%	50,00%
20	4	2	7	2	0	2	50,00%	0,00%	50,00%
Total	**88**	**43**	**92**	**22**	**20,5**	**42,5**			
Mean		**(48,86%)**					**24,19%**	**25,94%**	**50,13%**

* Descriptors assigned in the LISA database

Table VII
Quantitative relation between Keywords and Descriptors

		Keywords	Descriptors
INSPEC			
	Total:	730	775
	Mean:	4,6	4,9
CAB			
	Total:	841	1955
	Mean:	5,3	12,2
ISTA	Total:	724	646
	Mean:	4,5	4
LISA			
	Total:	724	780
	Mean:	4,5	4,9

Table VIII
Presence of Keywords in the INSPEC Descriptors

Source	Keywords	Descriptors
LOW presence		
Sigplan Notices Source: 2000, 35 (9): 23-33	1. Global register allocation 2. Graph coloring 3. Linear scan 4. Binpacking	1. GRAPH-COLOURING 2. OPTIMISING-COMPILERS 3. STORAGE-ALLOCATION
HIGH presence		
Theoretical Computer Science Source: 2002, 281 (1-2): 455-469	1. Random-transform 2. Discrete inverse problem 3. Discrete tomography 4. Contingency table 5. Computacional complexity 6. Polynomial-time algorithmic 7. NP-hard	1. COMPUTATIONAL-COMPLEXITY 2. DISCRETE-TRANSFORMS 3. INVERSE-PROBLEMS 4. RANDOM-TRANSFORM

Table IX
Presence of Keywords in the CAB ABASTRACT Descriptors

Source	Keywords	Descriptors
LOW presence		
Annual Review of Nutrition Source: 2001, 21: 283-295	1. LDL 2. Genetics 3. Diet 4. Subclass 5. Coronary disease	1. CARDIOVASCULAR-DISEASES 2. DIETARY-CARBOHYDRATE 3. DIETARY-FAT 4. DIETS 5. FOOD-INTAKE 6. GENES 7. HEART-DISEASES 8. LOW-DENSITY-LIPOPROTEIN 9. METABOLISM 10. REVIEWS 11. RISK 12. MAN
HIGH presence		
Water, Air & Soil Pollution Source: 2001, 132 (3-4): 215-231	1. PAHs 2. Hetero-PAHs 3. Soil 4. Biodegradation 5. Metabolites	1. Aromatic-hydrocarbons 2. Biodegradation 3. Composts 4. Metabolites 5. Polycyclic-hydrocarbons 6. Soil 7. Soil-pollution 8. Toxicity

Tabla X
Presence of Keywords in the ISTA Descriptors

Source	Keywords	Descriptors
LOW presence		
Online Information Review Source: 2002, 26 (2): 92-100	1. Online retrieval 2. Computing 3. Databases 4. Information Industry	1. Information-industry 2. Users 3.Information-professionals 4. History-of-information-science
HIGH presence		
Information Processing & Management Source: 2001, 37 (5): 661-675	1. Citation analysis 2. Computer science 3. Scholarly publishing 4. World Wide Web	1. Citation-analysis 2. Scholarly-Publising 3. Information-dissemination 4. Computer science

Tabla XI
Presence of Keywords in the LISA Descriptors

Source	Keywords	Descriptors
LOW presence		
Journal of Documentation Source: 2002, 58 (1): 49-65	1. Indexes 2. Information retrieval	1. Subject-indexing 2. Fiction 3. Consistency 4. Users 5. Library-staff 6. Public-libraries
HIGH presence		
Online Information Review Source: 2001, 25 (4): 257-266	1. Chemistry 2. Search engines 3. Hazardous materials 4. Internet	1. Online databases 2. Occupational health and safety 3. Chemistry 4. Hazardous materials 5. Internet 6. World Wide Web 7. Searching

Table XII
Presence of Keywords in the Descriptors

	Articles analysed	Num. of Keywords[1]	Num. of Descriptors [2]	Exact presence as % [3]	Normalizations As % [4]	Total [5]
CAB	160	841	1955	43,49	17,09	**60,58 %**
LISA*	160	724	780	23,00	19,52	**42,52 %**
INSPEC	160	730	775	11,34	30,28	**41,62 %**
ISTA*	160	724	646	20,51	17,38	**37,89 %**
Mean				**24,59**	**21,07**	**45,66%**

* The same articles were used
[1] Keywords provided by authors of the 160 artícles
[2] Total number of Descriptors assigned by the indexers to the 160 artícles
[3] Exact coincidence as % between the Keywords of the authors and the indexers' Descriptors
[4] Kws submitted to normalization process (e.g. ISTA: Text retrieval → INFORMATION-RETRIEVAL ; LISA: Text retrieval → ONLINE-INFORMATION-RETRIEVAL)
[5] Sum of total % of Exact + % Normalizations

ANNEX 1

Journals and years

DATABASE	JOURNALS	YEARS	Papers
INSPEC	Computer Methods and Programs in Biomedicine	2004, 2003, 2002, 2001, 2000	20
	Computerized Medical Imaging and Graphics	2004, 2003, 2002, 2001, 2000	20
	Computing	2004, 2003, 2001, 1999, 1997, 1995, 1993	20
	Performance Evaluation	2004, 2003, 2002, 2001, 2000	20
	Siam Journal on Computing	2004, 2003, 2001, 1999, 1997, 1995, 1993	20
	Sigplan Notices	2004, 2002, 2000, 1998	20
	Telematics and Informatics	2004, 2003 2002, 2001, 2000	20
	Theoretical Computer Science	2004, 2002, 2000, 1998	20
CAB	Agriculture and Human Values	2004, 2003, 2002, 2001, 2000	20
	Annual Review of Nutrition	2004, 2003, 2002, 2001, 2000	20
	Environmental Biology of Fishes	2004, 2003, 2002, 2001, 2000	20
	Environmental Geochemistry and Health	2004, 2003, 2002, 2001, 2000	20
	European Journal of Nutrition	2003, 2002, 2001, 1999	20
	Journal of Agricultural and Food Chemistry	2002, 2001, 2000, 1999, 1998	20
	Water, Air & Soil Pollution	2003, 2002, 2001, 2000, 1999	20
	Water Resources Management	2003, 2002, 2001, 2000, 1999	20
ISTA	Cataloging & Classification Quarterly	2001, 1999	20
	Information Processing & Management	2001, 2000	20
	International Journal of Information Management	2003, 2002, 2001, 1999	20
	Journal of Documentation	2002	20
	Library Acquisitions: Practice & Theory	1998, 1996, 1995, 1994	20
	Library Collections, Acquisitions & Technical Services	2001, 2000, 1999	20
	Online Information Review	2002, 2001, 2000	20
	The Electronic Library	2001, 2000	20
LISA	Cataloging & Classification Quarterly	2001, 1999	20*
	Information Processing & Management	2001, 2000	20*
	International Journal of Information Management	2003, 2002, 2001, 1999	20*
	Journal of Documentation	2002	20*
	Library Acquisitions: Practice & Theory	1998, 1996, 1995, 1994	20*
	Library Collections, Acquisitions & Technical Services	2001, 2000, 1999	20*
	Online Information Review	2002, 2001, 2000	20*
	The Electronic Library	2001, 2000	20*
		Total Articles	**640**

* These articles are the same as those used to compare the keywords with the descriptors in the ISTA database.

ANNEX 2

Document number of the articles in the databases

ISTA
2901030;2902351;2902353;2903830;2903831;2903832;2903833;3000404;3002690;3002691;3003662;3003663;3102709;3102710;3102711;3304224;3304243;3306036;3306037;3306050;3402758;3402977;3402980;3500289;3500846;3500848;3500867;3500951;3501025;3501200;3501226;3501382;3501467;3501468;3501474;3501592;3501619;3501682;3501691;3501699;3501728;3501801;3501877;3501898;3501905;3502023;3502120;3502478;3502794;3502908;3503771;3503815;3503903;3503975;3503980;3503996;3600172;3600251;3600311;3600621;3600683;3601288;3601327;3601404;3601452;3601614;3601760;3601989;3602017;3602030;3602063;3602136;3602198;3602201;3602684;3602753;3602857;3602893;3603195;3603372;3603452;3603519;3700246;3700304;3700417;3700489;3700517;3700521;3700659;3700780;3700797;3700878;3700950;3700951;3700952;3700961;3700962;3700980;3700981;3701060;3701222;3701232;3701395;3701399;3701680;3701711;3702039;3702254;3702411;3702503;3702707;3702710;3702741;3702782;3702785;3702822;3702864;3702985;3703029;3703057;3703168;3703170;3703172;3703175;3703308;3703416;3703892;3703899;3703925;3703972;3703979;3800039;3800102;3800127;3800149;3800518;3800615;3800619;3801024;3801224;3801274;3801337;3802655;3803495;3804250;3804254;3900022;EJ618329;EJ633003;EJ606816;EJ606818;EJ606784;EJ605358;EJ605357;EJ605356;EJ605364;EJ606787;EJ606786;EJ605363;EJ605361
LISA
2401070;2411817;2411819;2416838;2416839;2416840;2416841;2425816;2425817;2428709;2428710;2429830;2439531;2439532;2439533;2465669;2465670;2468652;2468653;2468655;3215418;3215419;3218188;3440438;3441301;3441302;3441303;3441306;3441308;3441309;3447417;3447418;3447419;3447420;3453825;3453826;3453830;3453831;3454179;3454180;3454181;3454182;3454184;3454186;3454252;3456134;3456136;3456137;3456147;3456149;3456150;3458220;3458221;3458222;3458354;3460468;3460469;3460470;3462008;3462010;3462011;3463939;3463940;3464228;3464229;3464230;3465568;3465569;3465570;3474088;3474403;3474549;3474560;3475603;3475627;3475628;3477231;3477232;3477819;3478109;3479626;3479627;3481557;3481770;3482475;3485433;3485434;3485435;3486256;3486617;3498336;3498342;3498371;3502016;3502017;3502018;3502019;3502020;3502021;3502029;3502395;3502396;3502411;3505606;3507641;3507642;3507643;3507644;3507645;3507646;3507647;3507652;3509910;3510112;3510113;3510114;3512619;3512620;3515829;3516447;3516448;3516449;3516451;3516452;3516453;3516454;3516455;3516456;3516457;3516458;3518083;3519759;3519957;3519960;3519961;3710476;3710477;3710478;3710511;3790658;3790659;3790905;3790906;3790907;3823633;3823634;3823635;3824075;3824076;3824077;3824260;3828112;3831955;3833090;3833092;3834294;3481558;3463941;3505605;3505608
INSPEC
4411592;4411598;4411601;4643668;4643669;4892490;4892492;4941455;4941458;4945306;4945311;4945313;5479784;5479785;5532651;5532655;5532657;5546819;5978044;5978056;5978062;6052758;6052759;6052763;6082448;6082450;6095133;6095135;6381161;6381163;6381165;6471658;6471660;6531104;6531106;6531335;6531336;6531338;6557340;6557342;6557343;6572942;6577788;6591136;6591137;6623752;6623754;6656175;6656177;6718086;6718090;6776200;6776205;6776231;6776232;6776236;6798985;6798986;6840222;6840226;6840280;6840282;6854974;6854975;6887969;6887971;6888173;6888175;6888178;6915242;6915244;6944702;6944703;6959979;6959980;6959982;6967310;6967318;7047932;7047934;7047937;7047938;7227286;7227288;7250446;7250449;7256753;7256754;7283503;7283504;7308850;7308854;7322750;7322751;7372753;7372754;7372765;7407093;7407095;7434750;7434753;7434754;7522303;7522305;7537268;7537272;7567650;7567653;7567659;7607407;7607409;7624203;7624207;7680369;7680371;7737123;7737125;7747619;7747620;7756546;7756548;7822815;7822816;7856994;7856996;7893564;7893565;7893567;7915552;7948440;7996894;7996897;8013494;8016528;8018227;8018228;8018234;8018238;8022764;8022766;8025083;8026854;8026857;8073372;8073375;8132714;8162652;8170681;8170682;8177749;8177750;8181573;8188014;8188017;8214533;8214534;8224592;8224594;8242702;8242704
CAB ABSTRACT
19981418346;19981418347;19981418348;19981418349;19990303175;19990303176;19990303177;19991903782;19991903783;19991903784;19991906904;19991910409;19991910410;19991910411;20000110662;20000311965;20000311966;20000311967;20000311968;20000402958;20001415197;20001415198;20001415199;20001415200;20001910155;20003003119;20003007753;20003007755;20003007757;20003007758;20003021330;20003033319;20003033322;20003033324;20003033327;20013001253;20013053731;20013095464;20013095470;20013095476;20013095478;20013122288;20013122291;20013122292;20013122297;20013132634;20013132895;20013132896;20013132903;20013132904;20013148266;20013148267;20013148268;20013148269;20013148534;20013148535;20013148536;20013148538;20013165074;20013165075;20013165077;20013165078;20013169370;20013169906;20013171014;20013171015;20013171018;20013171019;20023005865;20023011030;20023014736;20023014737;20023014739;20023014740;20023017851;20023017861;20023017885;20023089074;20023089075;20023089076;20023125277;20023125278;20023125279;20023125280;20023140813;20023152888;20023152889;20023152890;20023152891;20023158858;20023160381;20023160382;20023160383;20023167023;20023167024;20023167997;20023167998;20023167999;20023168000;20023198065;20023198282;20023198283;20023198284;20023198285;20033002379;20033002380;20033002381;20033002382;20033024560;20033030479;20033030480;20033030481;20033030482;20033037233;20033037235;20033037236;20033134940;20033134941;20033142542;20033142543;20033142544;20033142545;20033159928;20033159930;20033159931;20033159932;20033173145;20033173146;20033191836;20033191837;20033209800;20033209801;20033215939;20033215940;20033215941;20033215942;20043000064;20043000065;20043011891;20043011893;20043026335;20043026336;20043026337;20043026478;20043113950;20043113951;20043113952;20043113953;20043134370;20043155315;20043155316;20043155317;20043155318;20043179911;20043179912;20043179913;20043179915;20043185163;20043213155;20043213160

ANNEX 3

Relating to the keywords in the 'Instructions to authors' of the journals

	Journals	Author Instructions
INSPEC	Computer Methods and Programs in Biomedicine	3-6 key words for indexing purposes.
	Computerized Medical Imaging and Graphics	Key words: Enclose with each manuscript, at the end of the abstract, 5-10 key words. These terms should be relatively independent (coordinate index terms), and as a group should optimally characterize the paper.
	Computing	An AMS subject classification (primary, secondary) and suitable keywords and phrases should be given on the title page.
	Performance Evaluation	Please add one to five keywords to your article. Keywords are essential for the accessibility and retrievability of your article. Keywords assigned to articles will be assembled in a keyword index which will be printed in the last issue of each volume, and in cumulative indexes. In addition, it is planned to make keywords available on Internet. To maximize the consistency with which such keywords are assigned by different authors, the following guidelines have been drawn up…
	Siam Journal on Computing	1993: A list of key words must accompany all articles 1999 y 2004 : Key words and AMS subject classifications: List of key words and AMS subject classifications must accompany all articles.
	Sigplan Notices	No mencionada nada sobre keywords.
	Telematics and Informatics	2000-2001-2002 : No menciona nada sobre keywords. 2003-2004 : Immediately after the abstract, provide 3-5 keywords, avoiding general and plural terms and multiple concepts (avoid, for example, "and", "of"). Be sparing with abbreviations: only abbreviations firmly established in the field may be eligible. These keywords will be used for indexing purposes.
	Theoretical Computer Science	1998-2000 : Check to see that you have listed 3 to 5 keywords (to be placed under the abstract). Keywords are essential for the accessibility and retrievability of your article. 2002-2004 : Immediately after the abstract, provide a maximum of five keywords, using American spelling and avoiding general and plural terms and multiple concepts (avoid, for example, 'and', 'of'). Be sparing with abbreviations: only abbreviations firmly established in the field may be eligible. These keywords will be used for indexing purposes.
CAB ABSTRACT	Agriculture and Human Values	Submissions should include an abstract, not to exceed 250 words, a set of key words.
	Annual Review of Nutrition	No mencionada nada sobre keywords.
	Environmental Biology of Fishes	Finally, at bottom of the page the key words (no more than six) in lower case which should complement the title but not repeat words in it.
	Environmental Geochemistry and Health	… and the principal Conclusions. This should be followed by Keywords. (See below.)
	European Journal of Nutrition	Below the abstract place about 5 key words
	Journal of Agricultural and Food Chemistry	1998 : Provide four or five keywords to aid the reader in literature retrieval. The keywords are published immediately before the text for all papers and following the abstract (except for Rapid Communications). 2000 : Provide significant keywords to aid the reader in literature retrieval. The keywords are published immediately before the text for all papers and following the abstract. 2004 :Provide significant keywords to aid the reader in literature retrieval. The keywords are published immediately before the text, following the abstract.
	Water, Air & Soil Pollution	Please provide 5 to 10 key words or short phrases in alphabetical order.
	Water Resources Management	Key words supplied by the author should appear on a line following the abstract and will be used in a short index at the end of each volume of the journal. The key words selected should be comprehensive and subject specific. It is not necessary to list the subject area of the Journal's coverage as a key word. Six to 10 key words should be sufficient to cover the major subjects of a given paper, although more can always be supplied if the author deems it necessary. General terms should not appear as key words, as they have little use as information retrieval tools. Please, choose key words to be as specific as possible, and list the most specific first, proceeding to the most general last.
LISA ITSA	Cataloging & Classification Quarterly	The keywords be in the style of one of the major thesaureses [sic] … the terminology selected should be suitable for computer análisis.
	Information Processing & Management	Please also supply three to five keywords describing the main topics of the paper.
	International Journal of Information Management	Care should be taken to include up to five keywords suitable for indexing the article by computer analysis.
	Journal of Documentation	Up to six keywords should be included which encapsulate the principal subjects covered by the article.
	Library Acquisitions: Practice & Theory	Care should be taken to include up to five keywords suitable for indexing the article by computer analysis.
	Library Collections, Acquisitions & Technical Services	Up to six keywords should be included which encapsulate the principal subjects covered by the article.
	Online Information Review	Up to six keywords should be included wich encapsulate the principal subjects covered by the article.
	The Electronic Library	Five-six keywords that identify article content.

ANNEX 4

Detailed quantitative relation between Keywords and Descriptors

INSPEC database

Keywords	Num. of articles	Total Keywords	Descriptors	Num. of articles	Total Descriptors
1	0	0	1	4	4
2	7	14	2	19	38
3	36	108	3	34	102
4	49	196	4	22	88
5	35	175	5	23	115
6	16	96	6	23	138
7	11	77	7	15	105
8	2	16	8	4	32
9	2	18	9	10	90
10	1	10	10	4	40
20	1	20	11	1	11
			12	1	12
Total	**160**	**730**		**160**	**775**
Mean		**4,6**			**4,9**

CAB ABSTRACT database

Keywords	Num. of articles	Total Keywords	Descriptors	Num. of articles	Total Descriptors
1	0	0	1		
2	1	2	2		
3	12	36	3		
4	35	140	4	1	4
5	55	275	5	4	20
6	30	180	6	8	48
7	16	112	7	12	84
8	5	40	8	15	120
9	4	36	9	17	153
10	2	20	10	11	110
			11	13	143
			12	17	204
			13	16	208
			14	9	126
			15	5	75
			16	6	96
			17	3	51
			18	4	72
			19	4	76
			20	4	80
			22	2	44
			23	1	23
			24	3	72
			26	2	52
			28	1	28
			31	1	31
			35	1	35
Total	**160**	**841**		**160**	**1955**
Mean		**5,3**			**12,2**

ISTA database

	Keywords	Num. of articles	Total Keywords	Descriptors	Num. of articles	Total Descriptors
	1	0	0	1	0	0
	2	6	12	2	5	10
	3	25	75	3	7	21
	4	64	256	4	135	540
	5	38	190	5	4	20
	6	17	102	6	2	12
	7	4	28	7	5	35
	8	3	24	8	1	8
	10	1	10	10	1	10
	12	1	12			
	15	1	15			
Total		**160**	**724**		**160**	**646**
Mean			**4,5**			**4**

LISA database

	Keywords	Num. of articles	Total Keywords	Descriptors	Num.of articles	Total Descriptors
	1	0	0	1	1	1
	2	6	12	2	18	36
	3	25	75	3	27	81
	4	64	256	4	35	140
	5	38	190	5	24	120
	6	17	102	6	20	120
	7	4	28	7	13	91
	8	3	24	8	13	104
	10	1	10	9	6	54
	12	1	12	11	3	33
	15	1	15			
Total		**160**	**724**		**160**	**780**
Mean			**4,5**			**4,9**

ANNEX 5

Semantic relation between the keywords given by the authors and the indexers' descriptors

INSPEC	Num. PC	% PC used	% DE exact	% DE norm	Total %
Computer Methods Programs in Biomedicine	85	41,18%	12,94%	23,58%	35,42%
Computerized Medical Imaging and Graphics	108	51,85%	14,81%	40,45%	54,58%
Computing	82	51,22%	8,75%	37,32%	46,07%
Performance Evaluation	92	46,74%	9,67%	39,58%	48,00%
Siam Journal on Computing	80	37,50%	9,33%	25,50%	34,83%
Sigplan Notices	112	33,04%	5,48%	20,83%	26,30%
Telematics and Informatics	89	57,30%	23,17%	31,58%	54,13%
Theoretical Computer Science	82	39,02%	6,60%	23,43%	30,02%
Total	730	357,85%	90,74%	242,27%	329,36%
Mean	91,25	44,73%	11,34%	30,28%	**41,62%**

CAB ABSTRACT	Num. PC	% PC used	% DE exact	% DE norm	Total %
Agriculture and Human Values	117	66,67%	46,27%	20,21%	62,46%
Annual Review of Nutrition	99	28,28%	19,85%	6,73%	26,58%
Environmental Biology of Fishes	94	58,51%	28,21%	24,27%	52,49%
Environmental Geochemistry and Health	103	66,02%	53,18%	15,34%	66,43%
European Journal of Nutrition	99	68,69%	52,19%	11,51%	63,70%
Journal of Agricultural and Food Chemistry	116	64,66%	55,21%	15,96%	68,66%
Water, Air and Soil Pollution	104	69,23%	54,78%	19,85%	72,96%
Water Resources Management	109	57,80%	38,26%	22,82%	59,42%
Total	841	479,86%	347,95%	136,69%	472,70
Mean	105,1	59,98%	43,49%	17,09%	**60,58%**

ISTA	Num. PC	% PC used	% DE exact	% DE norm	Total %
Cataloging & Classification Quarterly	119	35,29%	22,43%	12,79%	35,22%
Electronic Library, The	94	44,68%	16,42%	18,92%	35,33%
Information Processing & Management	96	56,25%	36,42%	17,06%	53,48%
International Journal of Inform Management	79	41,77%	11,25%	31,04%	42,29%
Journal of Documentation	68	47,06%	9,92%	23,50%	33,42%
Library Acquisitions: Practice & Theory	88	42,05%	26,81%	13,29%	40,10%
Library Collec, Acquis & Tech Services	92	34,78%	22,85%	12,92%	35,77%
Online Information Review	88	30,68%	18,00%	9,54%	27,54%
Total	724	332,56%	164,10%	139,06%	303,15%
Mean	90,5	41,57%	20,51%	17,38%	**37,89%**

LISA	Num. PC	% PC used	% DE exact	% DE norm	Total %
Cataloging & Classification Quarterly	119	39,50%	20,58%	23,35%	43,93%
Electronic Library, The	94	46,81%	24,42%	18,54%	42,96%
Information Processing & Management	96	50,00%	25,75%	24,19%	49,94%
International Journal of Inform Management	79	40,51%	25,00%	10,00%	35,00%
Journal of Documentation	68	50,00%	26,67%	15,83%	42,50%
Library Acquisitions: Practice & Theory	88	48,86%	24,19%	25,94%	50,13%
Library Collec, Acquis & Tech Services	92	46,74%	13,55%	27,46%	41,01%
Online Information Review	88	37,50%	23,83%	10,83%	34,67%
Total	724	359,92%	183,99%	156,14%	340,14%
Mean	90,5	44,99%	23,00%	19,52%	**42,52%**