science softCon

Retrieval comparison from various text sources

K. H. Weber(1), G. K. Hartmann(2), H. Oberländer, M. L. Richards(3), A. K. Richter

Part 1 : Literature search on atmospheric ozone and water vapour from three different online-databases (Weber, K.-H., Hartmann, G.K., Oberländer, H.)

In context with an MPAE research project dealing with the validation of ozone (O_3) and water vapor (H_2O) data of the Earth's atmosphere the retrieved information from three bibliographic sources have been compared for the time period 1995-2000. Literature data from the online host STN International have been linked via the TOSCANA software to the numerical data. [1])

Bibliographic Sources :

Science Citation Index (SCI) is produced by the Institute for Scientific Information (U.S.A.). It contains about 16.7 million records from 1974 to date. It covers literature including cited references in natural sciences as well as medicine and humanities. Sources are more than 5.600 journals. SCI is available via the World Wide Web (http://www.isinet.com/products/citation/wos.html) as Web-of-Science (WoS) and also via STN International as the SCISEARCH database.

INSPEC is produced by the Institution of Electrical Engineers (IEE),UK , and FIZ Karlsruhe. It contains about 6 million records from 1969 to date. It covers literature in physics, electronics, electrical engineering, computers and information technology. Main sources are journals. INSPEC is available online via STN International.

ENERGY is produced in an international cooperation of International Energy Agency (IEA) member states within the program ETDE (Energy Technology Data Exchange). It contains about 3.7 million records from 1974 to date. It covers all fields of energy related research including environmental aspects. Sources are journals, books, conference papers, reports and patents. ENERGY is available online via STN International.

[1])For information on STN databases look at http://www.fiz-karlsruhe.de; for TOSCANA software see http://www.navicon.de

Results :

Number of documents found in the databases SCI (WoS), INSPEC, ENERGY (February 2000)

a) „atmospheric water vapor" (1995-2000)

search formulation	WoS/SCI	INSPEC	ENERGY
Water vapor or water vapour or H2O	26881	11758	4512
(water vapor or water vapour or H2O) and troposphere* [1)]	612	556	160
(water vapor or water vapour or H2O) and (troposphere* or stratosphere*)	958	817	232
(water vapor or water vapour or H2O) and (troposphere* or stratosphere* or atmosphere*)	3349	2453	991
(water vapor or water vapour or H2O) and (troposphere* or stratosphere* or mesosphere* or atmosphere*)	3364	2456	991

b) „atmospheric ozone" (1995-2000)

search formulation	WoS/SCI	INSPEC	ENERGY
Ozon* or O_3	11353	6179	5059
(Ozon* or O_3) and troposphere*	1575	862	652
(Ozon* or O_3) and (troposphere* or stratosphere*)	2968	1712	1113
(Ozon* or O_3) and (troposphere* or stratosphere* or atmosphere*)	4387	2565	2069

Both searches were executed in the following database fields :

INSPEC : Title, abstract, controlled terms (thesaurus keywords), supplementary terms
ENERGY : Title, abstract, controlled terms (thesaurus keywords)
SCI (WoS): Title, abstract, supplementary terms (author keywords), supplementary terms plus

Frequency distribution (%) by document types :
(Ozone search)

Document Type	SCI/WoS	INSPEC	ENERGY
Journal	100	91.3	40.5
Conference		17.4	33.6
Report or Report Article		1.2	26.4
Book or Book Article		<0.1	24.8
Miscellaneous			7.7
Numerical Data			6.0
Short Communication			3.6
Progress Report			3.1
Dissertation			2.8
Patent			< 1

document types may overlap thus leading to totals of > 100 %

Frequency distribution (%) by publication year :

Ozon search (Water vapour search)

Publication Year	SCI/WoS	INSPEC	ENERGY
1995	16.9 (16.8)	15.9 (16.8)	30.0 (29.0)
1996	19.0 (18.0)	22.0 (22.2)	26.6 (27.1)
1997	21.1 (21.1)	23.0 (22.4)	20.5 (22.1)
1998	19.7 (20.8)	22.0 (22.6)	13.4 (12.3)
1999	21.6 (21.4)	17.2 (16.0)	9.5 (9.5)

Comments and conclusions:

A relative simple search strategy has been applied because the results of the different databases should be comparable; thus only the so called basic index has been searched, covering the comparable database fields in each of the sources. For a more specific search one has to make use of database-specific fields.

Looking only at the total number of retrieved records, WoS (SCI resp.) shows the broadest coverage of literature concerning the investigated topic. This obviously results from the higher number of journals and a wider scope compared to INSPEC. Both SCI and INSPEC almost exclusively cover documents from scientific journals. ENERGY contains additional important information sources such as books, reports, Ph.D. thesis, and other so called „gray literature". Another database containing reports of governmental funded projects is NTIS (National Technical Information Service (U.S.A).

Concerning the number of publications per year,WoS and INSPEC show an increase from 1995 to 1996 and only little variation in the following years (the publications of 1999 are not yet completely stored in the databases).

The continuous decrease of the number of retrieved documents in ENERGY may be due to the fact that also the total number of all documents delivered to ENERGY has decreased over this period; the observed decrease thus is not linked to the applied topic.

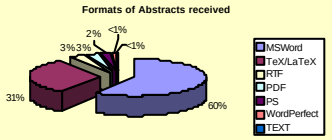
The database-specific enhancement of the search strategy (not shown here) resulting in a stronger relation to the data made available on the DUST-CD leads to questions concerning the usefulness of the available database-thesauri for searches in the field of atmospheric sciences. Continuous and careful enhancements and updates of the thesauri are recommended.

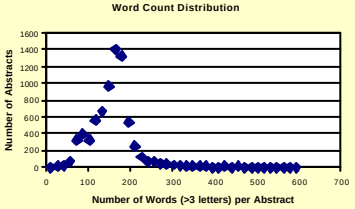
Part 2a, The EGS 2000 Conference Abstracts CD ROM
(Richards Copernicus Gesellschaft e.V.)

Abstracts for EGS 2000 in Nice have been processed at Copernicus Gesellschaft for reasonable uniformity of appearance. Authors could submit abstracts in practically any digital form. The diagram on the right summarises the formats actually used for the submitted abstracts.

The submitted files were converted to pdf file format for the EGS website and for the Abstracts CD-ROM. Unformatted text files were required for the SOMAccess processing, so all of the pdf files were subsequently converted to text. Some problems with the conversion to pdf could be identified more readily from the text files than the original or pdf versions.

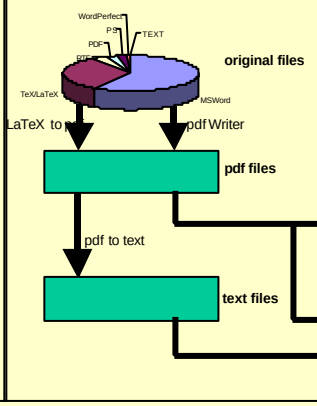
A by-product of this text analysis was the distribution of words used per abstract, illustrated in the second figure. For this purpose strings longer than 3 characters, including letters only, were counted as words. By this criterion 95% of the abstracts have a length of 218 words or less.





Part 2b, Processing EGS2000 Conference Abstracts for CD-ROM & SOMAccess

The following flow diagram summarises the work required to prepare the abstracts for the Abstracts CDROM and the SOMAccess processing.



Several problems were encountered in transforming from original file formats into pdf and then into plain text. The table lists the more quantifiable of these, in all less than 4% of the total. It does not include problems of converting the pdf files for the CD ROM.

Contact

Dr. Karl Heinz Weber, (Part 1)(1)
FIZ Karlsruhe
D-76012 Karlsruhe
Internet:www.fiz-karlsruhe.de

Prof. Dr. Gerd K. Hartmann, (PI) (2)
Max-Planck-Institut für Aernomie
Max-Planck-Straße 2
D-37191 Katlenburg-Lindau
www.linmpi.mpg.de

Dr. Michael Richards, (Part 2) (3)
Copernicus Gesellschaft e.V.
Max-Planck-Straße 13
D-37191 Katlenburg-Lindau
www.copernicus.org