

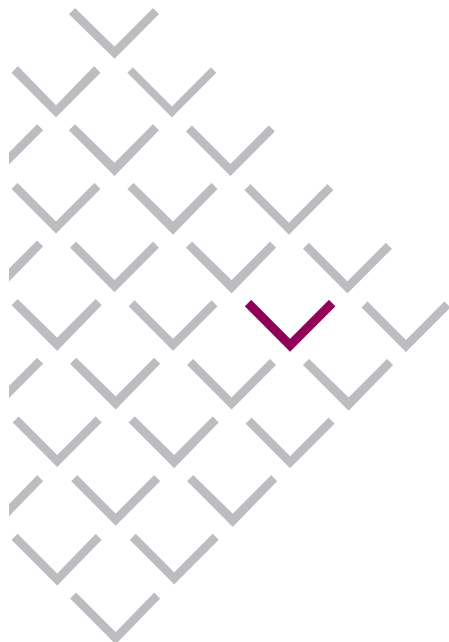


Chemical Abstracts Service
provides access to STN International
in North America

STNNotes

MAY 2003

No. 17 REVISED



ANALYZE and TABULATE Commands

Analysis of search results is more powerful than ever on STN® with the ANALYZE and TABULATE commands. You can view the analyzed terms not only as a one-dimensional list but also a two-dimensional table.

STNotes give you the in-depth technical details you need for efficient use of STN. For web access to STNotes, visit:
www.cas.org/ONLINE/STN/STNOTES/stnotescover.html.

Analysis of search results is easier and more powerful than ever on STN with the ANALYZE command for analysis of search results and the TABULATE command for correlating analyzed data.

The ANALYZE Command

You'll quickly find that ANALYZE is easy to use, with its clear prompts and intuitive defaults. With ANALYZE you can extract and analyze up to 50,000 terms from an answer set, even one that was created with a multifile search.

And you have many options for viewing the results. With the DISPLAY command you can view your analyzed terms as a one-dimensional list. In addition, if you extract terms from two or more fields, you can use the TABULATE command to create a 2-dimensional table of extracted terms.



To use the ANALYZE command

To use the ANALYZE command enter the file or files in which the L-number was created. Enter ANALYZE and answer the prompts for an L-number, answer numbers, and the field code(s).

```
=> ANALYZE
ENTER ANSWER SET OR ANALYZE L#
OR (L1):L1
ENTER ANSWER NUMBER OR RANGE
(1-):1-
ENTER DISPLAY CODE (FILEDEFAULT)
OR ?:PA
L2 ANALYZE L1 1- PA : 24 TERMS
```

You can use a single-file answer set or a multifile answer set, and L-numbers created by ANALYZE are also permitted. Only one L-number is allowed. The default is the last L-number created. To see a list of valid field codes for ANALYZE, enter HELP EFIELDS at an arrow prompt in the file or check the file's database summary sheet online in the STNGUIDE file, in print, or from the CAS web page (www.cas.org). You may enter up to five fields, separated by a space or comma.

When the ANALYZE process has completed, an L-number is created. The resulting L-number may be used in both the DISPLAY and TABULATE commands.

For online information on ANALYZE, enter HELP ANALYZE at an arrow prompt.

Display options after ANALYZE

Enter the DISPLAY command to view the analyzed results as a one-dimensional list of terms. The DISPLAY command prompts you for the L-number and the terms to display. The default is to display the top 10 most frequently occurring terms from the last created L-number.

```
=> DISPLAY
ENTER (L3), L# OR ?:L3
DISPLAY (TOP 10), ENTIRE OR ??:

Specify the number of terms you wish to display.
Valid responses are:

ENTIRE      display all of the terms
TOP n       where n refers to a number of terms (i.e. TOP 10)
OGT n       display terms with occurrence counts greater than 'n'
DGT n       display terms with document counts greater than 'n'
%GT n       display terms with percentage counts greater than 'n'
PGT n       display terms with percentage counts greater than 'n'
m-n         display the terms specified (e.g. 1-5)
DISPLAY (TOP 10), ENTIRE OR ?:
```

In addition, you have other DISPLAY options. You are not prompted for this information, but may enter it on the command line.

DISPLAY option	Example
• Display only terms containing a specified character string (in quotes)	=> DISPLAY L1 1- WITH "US"
• Display only terms that do not contain a specified character string	=> DISPLAY L1 1- NOT "US"
• Display the answer number with the terms	=> DISPLAY L1 1- ANS
• Change the default sort order to the alphabetical order and display terms	=> DISPLAY L1 1- ALPHA
• Change the sort order to the order of the frequency of occurrence in documents, and display terms	=> DISPLAY OCC 1-
• Display the terms in the delimited format for downloading	=> DISPLAY 1- DELIMITED

Editing and combining of extracted terms is possible with the EDIT command. Editing is especially useful when you have variations in the way a term, for example, a patent assignee, may appear in the database. With EDIT you can combine terms for the same patent assignees. Enter HELP EDIT at an arrow prompt for information on how to use EDIT.

Example:

In the last several years, what companies or organizations have applied for patents in the area of cancer vaccine development?

```

=> FILE WPINDEX

=> S CANCER(S)VACCINE? AND PRY>=2000
L1          736 CANCER(S)VACCINE? AND PRY>=2000

=> ANALYZE
ENTER ANSWER SET OR ANALYZE L# OR (L1):L1
ENTER ANSWER NUMBER OR RANGE (1-):1-
ENTER DISPLAY CODE (TI) OR ?:PA
L2          ANALYZE L1 1- PA :      985 TERMS

=> D
L2          ANALYZE L1 1- PA :      985 TERMS

TERM #      # OCC      # DOC      % DOC PA
-----
1           61         61         8.29 SMITHKLINE BEECHAM CORP
2           55         55         7.47 CORIXA CORP
3           50         50         6.79 SMITHKLINE BEECHAM PLC
4           32         32         4.35 MERCK PATENT GMBH
5           29         29         3.94 CURAGEN CORP
6           26         26         3.53 HUMAN GENOME SCI INC
7           22         22         2.99 DIADEXUS INC
8           14         14         1.90 GENZYME CORP
9           14         14         1.90 US DEPT HEALTH & HUMAN SERVICES
10          13         13         1.77 SMITHKLINE BEECHAM BIOLOGICALS

=> D TOP20
L2          ANALYZE L1 1- PA :      985 TERMS

TERM #      # OCC      # DOC      % DOC PA
-----
1           61         61         8.29 SMITHKLINE BEECHAM CORP
2           55         55         7.47 CORIXA CORP
3           50         50         6.79 SMITHKLINE BEECHAM PLC
4           32         32         4.35 MERCK PATENT GMBH
5           29         29         3.94 CURAGEN CORP
6           26         26         3.53 HUMAN GENOME SCI INC
7           22         22         2.99 DIADEXUS INC
8           14         14         1.90 GENZYME CORP
9           14         14         1.90 US DEPT HEALTH & HUMAN SERVICES
10          13         13         1.77 SMITHKLINE BEECHAM BIOLOGICALS
11          12         12         1.63 AGENSYS INC
12          11         11         1.49 SLOAN KETTERING INST CANCER RES
13          11         11         1.49 SUN Y
14          10         10         1.36 UNIV CALIFORNIA
15          10         10         1.36 XU J
16           9           9         1.22 HYSEQ INC
17           9           9         1.22 RECIPON H E
18           9           9         1.22 RUBEN S M
19           8           8         1.09 CAFFERKEY R
20           8           8         1.09 DILLON D C
21           8           8         1.09 LIU C
22           8           8         1.09 LUDWIG INST CANCER RES
23           8           8         1.09 MACINA R A
  
```

Search in selected file(s).

Enter **ANALYZE**, the L-number of the answer set (**L1**), the number of answers (**1-**), and the field code to analyze (**PA**).

A new L-number (**L2**) is created.

To display the 10 most frequently occurring terms, enter **D**.

To display the top 20 most frequently occurring terms, enter **D TOP20**.



```
=> D ALPHA 1-10
L2          ANALYZE L1 1- PA :      985 TERMS

TERM #      # OCC # DOC % DOC PA
-----
1           1     1   0.14 A+ SCI INVEST AB
2           1     1   0.14 ABU-THREIDEH J
3           1     1   0.14 ACAD ZIEKENHUIS LEIDEN
4           1     1   0.14 ADEREM A
5           1     1   0.14 ADJEI A L
6           1     1   0.14 ADVANCED POLYMER SYSTEMS INC
7           2     2   0.27 AEOMICA INC
8           1     1   0.14 AEROPHARM TECHNOLOGY INC
9           4     4   0.54 AFAR D E H
10          2     2   0.27 AFAR D E
```

```
=> EDIT
ENTER L#, E#, E# RANGE, OR (ALL):L2
ENTER (CHANGE), COMBINE, OR TITLE:COMBINE
ENTER PREFERRED TERM NUMBER OR (?):9
PREFERRED TERM:  AFAR D E H/PA
ENTER EQUIVALENT TERM NUMBERS OR (END):10
EQUIVALENT TERM:  AFAR D E/PA
ENTER EQUIVALENT TERM NUMBERS OR (END):END
APPLY CHANGES? (Y)/N:Y
TERMS COMBINED
```

```
=> D 1-10
L2          ANALYZE L1 1- PA :      985 TERMS
           (AFTER EDITS :      984 TERMS)

TERM #      # OCC # DOC % DOC PA
-----
1           1     1   0.14 A+ SCI INVEST AB
2           1     1   0.14 ABU-THREIDEH J
3           1     1   0.14 ACAD ZIEKENHUIS LEIDEN
4           1     1   0.14 ADEREM A
5           1     1   0.14 ADJEI A L
6           1     1   0.14 ADVANCED POLYMER SYSTEMS INC
7           2     2   0.27 AEOMICA INC
8           1     1   0.14 AEROPHARM TECHNOLOGY INC
9@          6     6   0.82 AFAR D E H
10          1     1   0.14 AGARWAL P
```

@ INDICATES TERM AFFECTED BY MOST RECENT EDITS

```
=> DISPLAY L2 OCC TOP 5
L2          ANALYZE L1 1- PA :      985 TERMS
           (AFTER EDITS :      984 TERMS)

TERM #      # OCC # DOC % DOC PA
-----
1           61    61   8.29 SMITHKLINE BEECHAM CORP
2           55    55   7.47 CORIXA CORP
3           50    50   6.79 SMITHKLINE BEECHAM PLC
4           32    32   4.35 MERCK PATENT GMBH
5           29    29   3.94 CURAGEN CORP
```

Reset the sort order and display the patent assignees in alphabetical order (D ALPHA).

Use EDIT to combine terms for the same patent assignee.

Display the first 10 terms in the changed order.

To reset the sort order from alphabetical to frequency of occurrence, enter OCC on the DISPLAY command line.

Additional ANALYZE options

You may use ANALYZE to extract terms only if they meet certain criteria. To extract terms that contain a 1-20 character string, add WITH followed by the character string in quotes. To extract all terms that do not contain a 1-20 character string, add NOT followed by the character string in quotes. To extract only the first n characters of each term, add LEN n to ANALYZE. LEN may be used with WITH or with NOT in one ANALYZE command. WITH and NOT may not be used in the same ANALYZE command.

ANALYZE may also be used to further refine an L-number with terms previously extracted by another ANALYZE command. To use ANALYZE for refinement, enter ANALYZE, the L-number, and the term numbers or range criteria you wish to use for analysis. The range criteria determine which terms are eligible for selection, and only one may be specified per ANALYZE command.

Range criteria applicable only for ANALYZE L-numbers are the first n terms based on current ordering, e.g., TOP 10; the terms with occurrence counts greater than n, e.g., OGT 100; the terms with document counts greater than n, e.g., DGT 15; and the terms with % of document counts greater than n, e.g., PGT 2.5.

To ANALYZE terms based on character strings

```
=> ANALYZE L5 PN 1- WITH "JP"           to extract only
                                           Japanese patents

=> ANALYZE L5 PN 1- NOT "JP"           to extract all
                                           patents except
                                           Japanese patents
```

To ANALYZE terms in an L-number created by ANALYZE

```
L2          ANALYZE L1 1- PA :          33 TERMS

=> ANALYZE L2 OGT 4
L6          ANALYZE L2 OGT 1 :          11 TERMS

=> D

L6          ANALYZE L2 OGT 1 :          11 TERMS

TERM #      # OCC  # DOC  % DOC PA
-----
      1         11    11   17.74 MINNESOTA MINING AND MFG. CO., USA
      2          5     5    8.06 XEROX CORP., USA
*****      END OF L6      *****
```

Fee for ANALYZE

When ANALYZE is used to create an L-number from display records, there is a fee based on the number of records analyzed. Enter HELP ANALYZE at an arrow prompt for cost information.

When ANALYZE is used to refine an ANALYZE L-number, there is no additional fee.



The TABULATE command

You can take your analysis of search results (the ANALYZE L-number) a step further with the TABULATE command - unique to STN. TABULATE gives you a two-dimensional display of terms extracted with ANALYZE from two display fields.

TABULATE is used mainly in patent and engineering databases and in LCA and LWPI. All extraction fields except patent number, application number, priority number, CAS Registry Number®, accession number, Basic Index, and similar fields, may be used.

To use the TABULATE command

To use the TABULATE command, enter TABULATE and answer the prompts.

You are prompted for information on the primary and secondary terms to display, i.e., the display field code, terms to display, and the sort order and direction. You can choose to sort terms by the frequency of their occurrence in documents (DOC) or in alphabetical order (ALPHA). Sort direction may be A for ascending (from lowest to highest for numbers, A to Z for alphabetic terms) or D for descending (highest to lowest for numbers, Z to A for alphabetic terms). The default sort direction is D for numbers and A for alphabetic terms.

You can choose from two display formats. The grid format is a two-dimensional table with the terms for the primary display code listed on the vertical axis and the terms for the secondary display code listed on the horizontal axis. In the default format, the terms are listed in separate columns. The first column lists the terms with the primary field code. The second column lists the terms with the secondary field code.

In addition, you have other TABULATE options. You are not prompted for this information, but may enter it on the command line.

TABULATE option	Example
• Display only terms containing a specified character string (in quotes)	=> TABULATE WITH "US"
• Display only terms that do not contain a specified character string	=> TABULATE NOT "US"
• Display the answer number with the terms	=> TABULATE ANS
• Display the terms in the delimited format for downloading	=> TABULATE DELIMITED

You can also enter the unprompted information for TABULATE in response to the following prompt:

```
REFORMAT USING SAME DISPLAY FIELDS? (N), Y, OR ?:TABULATE DELIMITED
```

Fee for TABULATE

Enter HELP TABULATE at an arrow prompt for online information on the command and the fee.

There is a fee for the use of each TABULATE command. However, within each TABULATE command you can reformat the display using the same display fields for only one fee per command. For example, you can display the same term in the default and the grid format within the same TABULATE command for only one fee.

Example of TABULATE command

Who are the top 10 researchers in the area of nanoelectronics and when did they publish their findings?

```

=> FILE INSPEC COMPENDEX

=> S (NANOELECTRON? OR NANO? AND SINGLE ELECTRON) AND PY>1993
L1      2038 (NANOELECTRON? OR NANO? AND SINGLE ELECTRON) AND PY>1993

=> DUP REM L1
PROCESSING COMPLETED FOR L1
L2      1604 DUP REM L1 (434 DUPLICATES REMOVED)
        ANSWERS '1-1410' FROM FILE INSPEC
        ANSWERS '1411-1604' FROM FILE COMPENDEX

=> ANALYZE L2 1- AU PY
L3      ANALYZE L2 1- AU PY :      3845 TERMS

=> TABULATE
ENTER ANALYZE OR SELECT L# (L3):L3
DISPLAY AS GRID FORMAT (N), Y, OR ?:Y
ENTER PRIMARY DISPLAY CODE OR (?):AU
ENTER SECONDARY DISPLAY CODE OR (?):PY
DISPLAY PRIMARY (TOP 10), ENTIRE OR ?:.
PRIMARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:DOC
PRIMARY SORT DIRECTION (DEFAULT), A, D, OR ?:D
SECONDARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:DOC
SECONDARY SORT DIRECTION (DEFAULT), A, D, OR ?:D
A FEE WILL BE CHARGED. PROCEED? (Y), N, OR ?:Y
L3      ANALYZE L2 1- AU PY :      3845 TERMS

```

AU	PY									
	2002	2001	2000	1999	1998	1997	1996	1995	2003	1994
AHMED, H.	5	6	3	7	8	2	2	2	0	0
MATSUMOTO, K.	1	2	5	4	5	4	5	2	0	0
HASEGAWA, H.	4	5	3	5	2	0	1	0	2	0
POROD, W.	4	3	3	2	2	3	0	0	2	0
ODA, S.	2	3	10	0	1	2	0	0	0	0
SAMUELSON, L.	4	3	1	3	2	1	0	1	0	0
KASAI, S.	4	4	2	1	1	0	0	0	2	0
TAKAHASHI, Y.	4	1	2	2	3	1	0	1	0	0
BANDYOPADHYAY, S.	2	1	0	0	2	3	3	0	0	1
BLICK, R.H.	2	4	2	2	0	0	0	1	1	0
HIRAMOTO, T.	1	2	3	4	1	0	1	0	0	0
NAGASE, M.	1	0	2	4	1	2	0	1	1	0

Search in selected files.

ANALYZE two fields (AU and PY).

Enter TABULATE.

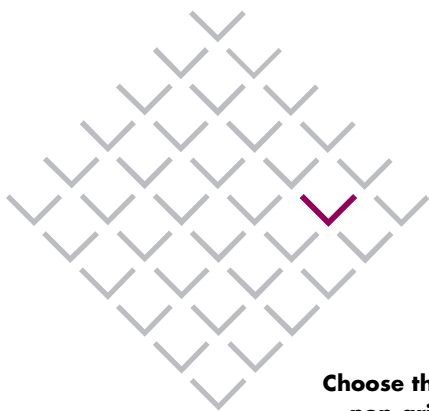
Enter the ANALYZE L-number.

Choose the GRID format.

Enter the primary and the secondary display code and sort requirements.

Enter Y to proceed.

The top 10 authors are listed on the vertical axis. The publication years are listed on the horizontal axis. The document count is shown at each intersection of AU and PY.



Choose the non-grid format.

Enter the display requirements.

The top 10 authors are listed in the AU column. For each author, the publication years are listed in the PY column. For each term, the number and percent of documents is displayed.

Enter TABULATE DELIMITED to reformat the same fields.

Enter the display requirements.

Terms are displayed in the delimited format for downloading.

Enter N to exit the TABULATE command without reformatting.

```
REFORMAT USING SAME DISPLAY FIELDS? (N), Y, OR ?:Y
DISPLAY AS GRID FORMAT (N), Y, OR ?:N
EXCHANGE PRIMARY AND SECONDARY DISPLAY FIELDS (N), Y, OR ?:.
DISPLAY PRIMARY (TOP 10), ENTIRE OR ?:.
DISPLAY SECONDARY (TOP 10), ENTIRE OR ?:.
PRIMARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:.
PRIMARY SORT DIRECTION (DEFAULT), A, D, OR ?:.
SECONDARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:.
SECONDARY SORT DIRECTION (DEFAULT), A, D, OR ?:.
L3          ANALYZE L2 1- AU PY :    3845 TERMS
```

TERM #	# DOC	% DOC	AU	PY
10	35	2.18	AHMED, H.	
	8	0.50	—	1998
	7	0.44	—	1999
	6	0.37	—	2001
	5	0.31	—	2002
	3	0.19	—	2000
	2	0.12	—	1995
	2	0.12	—	1996
	2	0.12	—	1997
12	28	1.75	MATSUMOTO, K.	
	5	0.31	—	1996
	5	0.31	—	1998
	5	0.31	—	2000
	4	0.25	—	1997
	4	0.25	—	1999
	2	0.12	—	1995
	2	0.12	—	2001
	1	0.06	—	2002
13	22	1.37	HASEGAWA, H.	
	5	0.31	—	1999
	5	0.31	—	2001


```
REFORMAT USING SAME DISPLAY FIELDS? (N), Y, OR ?:TABULATE DELIMITED
DISPLAY AS GRID FORMAT (N), Y, OR ?:N
EXCHANGE PRIMARY AND SECONDARY DISPLAY FIELDS (N), Y, OR ?:.
DISPLAY PRIMARY (TOP 10), ENTIRE OR ?:.
DISPLAY SECONDARY (TOP 10), ENTIRE OR ?:.
PRIMARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:.
PRIMARY SORT DIRECTION (DEFAULT), A, D, OR ?:.
SECONDARY SORT ORDER (CURRENT), DOC, ALPHA, OR ?:.
SECONDARY SORT DIRECTION (DEFAULT), A, D, OR ?:.
L3          ANALYZE L2 1- AU PY :    3845 TERMS
```

```
8;0.50;AHMED, H.;1998
7;0.44;AHMED, H.;1999
6;0.37;AHMED, H.;2001
5;0.31;AHMED, H.;2002
3;0.19;AHMED, H.;2000
2;0.12;AHMED, H.;1995
2;0.12;AHMED, H.;1996
2;0.12;AHMED, H.;1997
5;0.31;MATSUMOTO, K.;1996
.
```

```
REFORMAT USING SAME DISPLAY FIELDS? (N), Y, OR ?:N
```