

Abstract for

Taming the Beast: A Software Tool for Managing Field Verbosity using SAS Macros and SCL

**Betty Smith, Computer Programmer
Bureau of Business Research
University of Texas at Austin**

A software tool we call Key Word Search has been successfully employed in the processing of the data in a large manufacturers database. The tool uses SCL and the SAS Macro Facility to search a 200 character product description field for common words defined by user input. This paper describes how we at the Bureau of Business Research are currently using the tool for a number of tasks. The tool is easily adaptable to almost any application; but, will probably find a home in organizations with a need to search and manipulate large databases

**Taming the Beast: A Software Tool for Managing Field
Verbosity using SAS/Macro Facility™ and SAS/SCL™**

Management of a large database has spurred many unique problems of which field verbosity is one of the hardest to handle. When a field is too wordy and free form in its expression, it becomes extremely difficult to automate various tasks, such as ease in access, extraction, and processing of the data. Such is the case with the product description field in the database used for publication of the Directory of Texas Manufacturers. This field provides detailed product descriptions which reflect more accurately a company's manufacturing activities than any abbreviated or restrictively worded product description; however, it is very difficult to manage as a part of a viable electronic database. Electronic access to and indexing of this field have proven to be very difficult tasks. Until the creation of a software tool we call Key Word Search using SCL and the Macro facility, many of the tasks required more human processing than computer processing which was both inefficient and expensive. The following paper describes the tool and how we are currently using it with our manufacturers' database.

Product Description Field Defined

The product description field is 200 characters in length with a somewhat free form styling in its wording. For example here are few examples from the Directory of Texas Manufacturers for the same SIC¹ code.

3589 Equipment for treating water & wastewater
3589 Water and wastewater treatment plant equipment
3589 Wastewater treatment equipment
3589 Water treating equipment and assembly of
wastewater treating equipment
3589 Wastewater treatment equipment, aeration
equipment for catfish ponds²

While each of these descriptions describe the same product classification code, they do are not necessarily describe the same product. A catfish farmer, for example using the directory, would only be interested in the last description. If the field descriptions were uniform, the directory would be virtually useless to the catfish farmer who would have to call every manufacturer of water treating systems to find one that makes equipment only for catfish ponds; and if a viable index cannot be created from such unwieldy descriptions, he might still have a problem identifying under which SIC code the product is listed. This field is also used to index the Directory and to query the database.

What the Key Word Search Tool Is and How It Works:

The Key Word Search Tool is an integrated SCL and SAS program which uses macro variables to search the product description field for the occurrence of user defined common words, extracts that information, and presents the result in specifically designed formats for use in various applications.

The Key Word Search begins in an AF program (in another variation in FSEDIT) where the user is asked to define up to four key words (example 1.1). When the user exits the screen, an SCL program takes over which feeds the key words into an array called Key (example 1.2). Array Key is then polled to determine how many key words exist and the total is put in the variable called Pass. Array Key is then sorted in descending order using the sort function so that the blank key words are the last elements in the array. The elements of Array Key and Pass are then fed into macro variables &key1, &key2, &key3, &key4, and &ps (example 1.3) using the symput function.

Pass is then checked to see if it is greater than zero (meaning, of course, that Array Key is empty if Pass is equal to zero). If Pass is greater than zero then a submit block is called to begin the extraction process.

The extraction process (example 1.4) involves linking to the disk which stores the product description dataset called Retrieve.Sic. The macro variables which contain the key words and the key word count (&key1-&key4 and &ps) are

then imported into SAS variables Ky1-Ky4 and Ps. The product description (Sictext) and the product SIC code (Siccode) are fed into variables Sictn and Sictx. We search the product description field using the index function to locate the occurrences of the key words (notice the if-then-do blocks in the source program example 1.4).

Please enter Keywords below:

Keyword #1: _____
Keyword #2: _____
Keyword #3: _____
Keyword #4: _____

Ex. 1.1

```
array key(4) key1-key4;  
pass=0;
```

Ex. 1.2

```
do i = 1 to 4;  
  if key(i) ne ' ' then pass+1;  
end;  
call symput('ps',pass);  
rc=asort(key,'d');  
call symput('key1',key(1));  
call symput('key2',key(2));  
call symput('key3',key(3));  
call symput('key4',key(4));
```

Ex. 1.3

```

if pass > 0 then do;
control always;
alarm;
control asis;
submit immediate;
cms link brae775 192 394 rr dtm;
cms acc 394 q;
cms fi retrieve disk d d q;
data sichome.data(keep= id select ky1-ky4
sictx sictx);
length select $1 sictx $200 sictx 4;
retain ps1 0;
set retrieve.sic;
ps="&ps"*1;
ky1=trim(left("&key1"));
ky2=trim(left("&key2"));
ky3=trim(left("&key3"));
ky4=trim(left("&key4"));
sictx=sictext;sictx=siccode;
if ps = 1 then do;
indx1=index(upcase(sictext),upcase(ky1));
if indx1 > 0 then output;
end;
if ps = 2 then do;
indx1=index(upcase(sictext),upcase(ky1));
indx2=index(upcase(sictext),upcase(ky2));
if indx1 > 0 and indx2 > 0 then output;
end;
if ps = 3 then do;
indx1=index(upcase(sictext),upcase(ky1));
indx2=index(upcase(sictext),upcase(ky2));
indx3=index(upcase(sictext),upcase(ky3));
if indx1 > 0 and indx2 > 0 and indx3 > 0 then
output;
end;
if ps = 4 then do;
indx1=index(upcase(sictext),upcase(ky1));
indx2=index(upcase(sictext),upcase(ky2));
indx3=index(upcase(sictext),upcase(ky3));
indx4=index(upcase(sictext),upcase(ky4));
if indx1 > 0 and indx2 > 0 and indx3 > 0 and
indx4 > 0 then output;
end;
ps1=0;
run;

```

Ex. 1.4

**Variations on a Theme: How the Tool is Used in Our
Applications.**

--SIC coding of product descriptions:

First, I will describe the problem as it existed before the implementation of the Key Word Search Tool for entry operators assigning SIC codes product descriptions given to us by manufacturers:

In order to code a product description, the entry operator had primarily two sources: one, was the Standard Industrial Classification Manual 1987³ which is the standard for classification of products and services for industry in the U. S; the second source the entry operator had was a thousand page tome created from the SIC manual with very little additional input from the manufacturers' database.

Take for example the following:

A company lists as its product: Sandblasting equipment. This product can be coded as either 3559 or 3569. If you look up 'Sandblasting equipment' in the index of the SIC manual you will not find it, because both SIC codes 3559 and 3569 are somewhat general in how the book describes the codes (examples 2.1 & 2.2)³.

3559 Special Industry Machinery, Not Elsewhere Classified

Establishments primarily engaged in manufacturing special industry machinery, not else where classified, such as smelting and refining equipment, cement making, clayworking, cotton ginning, glass making, hat making, incandescent lamp making, leather working, paint making, rubber working, cigar and cigarette making, tobacco working, shoe making, and stone working machinery, and industrial sewing machines, and automotive maintenance machinery and equipment.

Ammunition and explosives
machinery
Anodizing equipment (except
mill lines)
Bag sewing and closing
machines (industrial
sewing machines)
.
.
.
Rubber products machinery
Rubber working machinery
Scouring machines tannery
Semiconductor manufacturing
machinery

Desalination equipment
Die and hub cutting equipment
(jewelry manufacturing)
Drying kilns, lumber
Electric screening equipment
Electron tube making machinery
Electroplating machinery
.
.
.
Wheel mounting and balancing
equipment, automotive
Wood drying kilns
Zipper making machinery

Ex. 2.1

3569 General Industrial Machinery and Equipment, Not Elsewhere Classified

Establishments primarily engaged in manufacturing machinery, equipment, and components for general industrial use, and for which no special classification is provided. Machine shops primarily engaged in producing machine and equipment parts, usually on a job or order basis, are classified in Industry 3599.

Altitude testing chambers
Baling machines for scrap
metal, paper, and
similar products
Blast cleaning equipment,
dustless; except
metalworking
Brake burnishing and wash-
ing machines
.
.
.
Driers and reel, firehose
Filter elements, fluid:
hydraulic line
Filters, fluid, general
line industrial:
except internal
combustion engines

Jacks, hydraulic: for general
use
Lubricating systems,
centralized
Lubrication equipment,
industrial
Lubrication machinery,
automatic
.
.
.
Robots for general industrial
use
Screening and sifting machines
Screws, jack
Separators for steam, gas,
and air (machinery)

Ex. 2.2

If 'Sandblasting equipment' is not listed under these codes then why was it assigned to these codes? The answer is very simple because the SIC manual has been interpreted, through its descriptive paragraphs, to mean that 'Sandblasting equipment' may go in either code depending on whether that equipment was designed for special industrial purposes or general industrial purposes (read the paragraphs in examples 2.1 & 2.2). But, the product description, 'Sandblasting equipment', received from the manufacturer, may not have any qualifying adjectives to indicate which of two codes is appropriate. You can see how frustrating it is to assign SIC codes.

When Key Word Search is employed, the entry operator can now view and code by example from actual entries existing in the product dataset (examples 2.3-2.5). As you can see, the obvious decision is to code this product description as 3569 since no other qualifiers are included in the original description.

The average run time for this application is about two minutes. The average time it takes for an entry operator to look up information can be five minutes or more depending on the degree of difficulty in finding the SIC code. Although interpretation is not entirely eliminated, the Key Word Search Tool has eliminated a lot guess work on the part the entry operator. Furthermore, if no specific code can be determined after the Key Word Search, the entry operator still has a better idea from viewing the examples of the type of questions to ask the manufacturer in a follow-up

Please enter Keywords below:

Keyword #1: sand
Keyword #2: blasting
Keyword #3: equipment
Keyword #4:

Ex. 2.3

+FSEDIT SICHOME.DATA-----Obs 1+

3 Total examples were generated by your descriptions.
1 Is the current example you are viewing.

Place an 'x' here if you wish to use this description.

Sic # 3559

Description: Metal-cleaning sandblasting equipment, including bl
asting nozzles & accessories

Paging instructions: F8 to view next example
F7 to view previous example
F3 to end this session

Ex. 2.4

+FSEDIT SICHOME.DATA-----Obs 2+

3 Total examples were generated by your descriptions.
2 Is the current example you are viewing.

Place an 'x' here if you wish to use this description.

Sic # 3569

Description: Sandblasting equipment

Paging instructions: F8 to view next example
F7 to view previous example
F3 to end this session

Ex. 2.5

phone call. She can even have the list on the screen when she calls.

--Retrieval by product description.

Information Services at the Bureau of Business Research provides a number of services including extractions from and queries of the manufacturers' database. Before Key Word Search was instituted, the customer would need to wait several days for Information Services to find the requested information.

First, the correct SIC code is determined for the product to be found by using both the SIC manual and the Directory index. Then the Directory is searched manually to find the product description or its equivalent based on the requirements requested by the customer.

When the Key Word Search was employed in this application, criteria including the key words were entered on an Fsed it screen (This is the only variation of key word which does not use SCL) and the information is extracted for viewing and/or reporting the information. In the following example:

A search was initiated for those companies, who list commercial offset printing as their product, located in Amarillo, TX (example 3.1). After the search is completed a report is made to the user (example 3.2) which gives a count of the companies and number of options. If the user chooses

DTM Search

Geographical Search Parameters:

Location City: Amarillo _____

Location County: _____

MSA Name: _____

Area Code: () _____

Mail Zip: _____ Location Zip: _____

Employee Range _____ to _____	and/or _____	Sales Range _____ to _____	and/or _____
Organization Code _____	and/or _____	Dist. Code _____	and/or _____
Year Established _____ to _____	and/or _____	Fax _____	and/or _____
Toll Free Numbers _____	xxxxxxxxxxxx		xxxxxxxxxxxx

Sic Text Search Parameters:

Sic Code: _____ (Enter 4, 3, or 2 digits)

Keywords:

Keyword # 1: commercial _____

Keyword # 2: offset _____

Keyword # 3: printing _____

Ex. 3.1

We have found a total of (2) matches and a total of (2) non-duplicated matches for the following search variables:

Location City: Amarillo _____

Location County: _____

MSA Name: _____

Area Code: () _____

Mail Zip: _____ Location Zip: _____

Employee Range _____ to _____	and/or and _____	Sales Range _____ to _____	and/or and _____
Organization Code _____	and/or and _____	Dist. Code _____	and/or and _____
Year Established _____ to _____	and/or and _____	Fax _____	and/or and _____
Toll Free Numbers _____	xxxxxxxxxxxx		xxxxxxxxxxxx

Sic Code: _____ (Enter 4, 3, or 2 digits)

Keywords:

Keyword # 1: commercial

Keyword # 2: offset

Keyword # 3: printing

Do you want to view this list? x

or

Bypass this list? _

Do you want to just print this list? _

Ex. 3.2

OBS	VIEW	PRT	FIRM
1	x		Southwestern Publications
2			Printco of Amarillo

Ex. 3.3

Id: 11416

Name: Southwestern Publications

Plt:

Addr: Box 10166

Amarillo

TX

791161166

Loc.: 100 Taylor

Phone: 8063723888

PE: Lee Bural

Title: plt mgr

City where plant is located: Amarillo

County plant is located: POTTER

Siccode Text

2752 Offset commercial printing, including newspapers,
mailers, magazines, & telephone directories

Ex. 3.4

to view the information (as we have in example 3.3), an Fsview screen is pulled up. If the user wishes to view the detailed information on a company(ies) (as we have in example 3.4) the information is displayed. The user may also choose to output the information by returning to the Fsview screen and putting an 'x' in the prt field.

--Automating the index:

Prior to the 1992 edition of the directory, indexing was done manually. After successfully using the tool in the two previous applications, it was decided to incorporate it in automating the indexing of the directory. The last time the index was completely updated was in 1989.

Since no true relational database exists for the index, updating the index with this new application was no less tedious than the old method. Virtually every entry in the product data set had to be reviewed not once, but many times (one product description may house multiple key words). For the future editions of the directory only those entries which do not have an index code will be dealt with.

To index a product, the key word(s) for the product is entered such as in example 4.1. The key word entered is 'rock'. An index code is automatically assigned to the record. For example, in this case the index code is R 001 005 (example 4.2), the first part of the index code is 'R' (based on the first letter of the key word), the second part of the code is '001' (if a pre-existing 'R' was in the

master index this part would be incremented), and the third part of the code is incremented when the SIC code changes. Notice that there are several fields in examples 4.2-4.10. The indexer can choose to discard a record from the index or can enter adjective and noun combinations which in conjunction with the key form the entry in the master index. What is entered in these fields depends on whether the key word is a noun or an adjective in the product description. The key words initiated in the search were primarily taken from the old index.

After completion of the initial key word search, non-coded entries in the product database were then extracted for assignment of key word(s), adjectives and noun (example 4.11). If more than one key word exists for the entry, the indexer could enter them in the key field separated by delimiters. The additional key word(s) were later pulled out creating new index records. This subsetted index was then appended to the master index and sorted. New index codes were then assigned to the master index.

After completion of this step, a compressed & cross indexed version of the index was created for use in typesetting the index (example 4.12). Once the index was typeset all like SIC codes under a specific key word(s) were put into a paragraph format similar to the following example:

Fence:

**frames for, metal; posts, fabricated
steel t-posts 3499**

During the updating of the database in the future, changes to the product description will cause the deletion of the record(s) in the master index relating to the id+SIC code link, thereby making it possible to extract and view the record(s). If a company eliminates the product, the record(s) in the master index will be deleted when the record is deleted from the product data set. If a company adds a product or the company is a new listing those records in the product index will, of course, not be represented in the master index until the index updating process starts.

When it is time to update the index, all records in the product information file that have no link to the master index will be extracted for viewing. At this time the subsetting records will be searched using the Key Word Search Tool automatically using the master index key words, and qualified for entries for the index. If key word(s) don't exist for some of the records then new key word(s) will be assigned using the programs used to create example 4.11. Once the new data has been massaged it will be appended to the master index, sorted, and reassigned new index codes.

Please enter Keywords below:

Keyword #1: rock _____
Keyword #2: _____
Keyword #3: _____
Keyword #4: _____

Ex. 4.1

+FSEdit SICH0.DATA-----Obs 12 Screen 1+

64 Total examples were generated by your descriptions.
12 Is the current example you are viewing.

Place an 'x' here if you do not wish to use this in the index.

Sic # 2899 Index Code: R 001 005
Description: Rock processing chemicals, concrete admixtures, & w
aterproofing compounds

Key: Rock

Adjective1: processing _____
Adjective2: _____
Noun: chemicals _____

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.2

64 Total examples were generated by your descriptions.
 13 Is the current example you are viewing.

Place an 'x' here if you do not wish to use this in the index.

Sic # 2951 Index Code: R 001 006
 Description: Limestone rock asphalt paving materials

Key: Rock

Adjective1: limestone
 Adjective2: asphalt paving
 Noun:

Paging instructions: F8 to view next example
 F7 to view previous example

Ex. 4.3

64 Total examples were generated by your descriptions.
 18 Is the current example you are viewing.

Place an 'x' here if you do not wish to use this in the index.

Sic # 3272 Index Code: R 001 010
 Description: Precast rock siding & bumper curbs

Key: Rock

Adjective1: precast
 Adjective2:
 Noun: siding

Paging instructions: F8 to view next example
 F7 to view previous example

Ex. 4.4

64 Total examples were generated by your descriptions.
19 Is the current example you are viewing.

_ Place an 'x' here if you do not wish to use this in the index.

Sic # 3295 Index Code: R 001 011

Description: Aggregates, sand, & landscaping rocks

Key: Rock

Adjective1: landscaping

Adjective2:

Noun:

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.5

64 Total examples were generated by your descriptions.
24 Is the current example you are viewing.

_ Place an 'x' here if you do not wish to use this in the index.

Sic # 3295 Index Code: R 001 011

Description: Crushed concrete road-base aggregate, ground rock paving aggregate

Key: Rock

Adjective1: ground

Adjective2: paving

Noun: aggregate

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.6

64 Total examples were generated by your descriptions.
29 Is the current example you are viewing.

_ Place an 'x' here if you do not wish to use this in the index.

Sic # 3451 Index Code: R 001 012

Description: Threaded rebar couplings, rebars, rock bolts, tie b
ars

Key: Rock

Adjective1: _____

Adjective2: _____

Noun: bolts _____

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.7

64 Total examples were generated by your descriptions.
31 Is the current example you are viewing.

_ Place an 'x' here if you do not wish to use this in the index.

Sic # 3531 Index Code: R 001 013

Description: Construction industry rock saws & plows for laying
cable

Key: Rock

Adjective1: construction _____

Adjective2: industry _____

Noun: saws _____

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.8

64 Total examples were generated by your descriptions.
 39 Is the current example you are viewing.

Place an 'x' here if you do not wish to use this in the index.

Sic # 3533 Index Code: R 001 015

Description: Oil field rock drill bits, including diamond drill
 bits

Key: Rock

Adjective1: oil field

Adjective2: drill

Noun: bits

Paging instructions: F8 to view next example
 F7 to view previous example

Ex. 4.9

64 Total examples were generated by your descriptions.
 41 Is the current example you are viewing.

Place an 'x' here if you do not wish to use this in the index.

Sic # 3559 Index Code: R 001 016

Description: Rock polishers

Key: Rock

Adjective1:

Adjective2:

Noun: polishers

Paging instructions: F8 to view next example
 F7 to view previous example

Ex. 4.10

+FSEdit WORK.SICHO-----Obs 1 Screen 1+

28931 Total examples were generated by your descriptions.
1 Is the current example you are viewing.

_ Place an 'x' here if you do not wish to use this in the index.

Sic # 3993

Index Code:

Description: Neon signs

Key: signs

Adjective1: neon

Adjective2:

Noun:

Paging instructions: F8 to view next example
F7 to view previous example

Ex. 4.11

OBS	INDX	KEY	SIGN	INDX1	SEEALSO	N
113	F001010	Fence	3446	posts, ornamental iron & steel fabrication of (1);		86
114	F001010	Fence	3446	posts, steel (1);		89
115	F001010	Fence	3446	posts, studded steel (1);		90
116	F001010	Fence	3446	steel (1);		112
117	F001011	Fence	3496	chain link (2);	(See also under Chain: Fence. . .)	13
118	F001011	Fence	3496	fittings, chain link (1);	(See also under Chain: Fence. . .)	31
119	F001011	Fence	3496	gates, chain link (3);	(See also under Chain: Fence. . .)	38
120	F001011	Fence	3496	interior partitions & fittings, chain link (1);	(See also under Chain: Fence. . .)	54
121	F001011	Fence	3496	parts, chain link (1);	(See also under Chain: Fence. . .)	74
122	F001012	Fence	3499	frames for, metal (1);		35
123	F001012	Fence	3499	posts, fabricated steel t-post (1);		84
124	F001013	Fence	3523	panels, steel (1);		72
125	F001014	Fence	3993	signs, metal (1);		101

232

Ex. 4.12

Conclusion:

The Key Word Search Tool has proven to be a great boon to improving the processes in the management manufacturers' database. It is hoped that other applications relating not only to the Directory of Texas Manufacturers' database, but other databases, will be able to employ the tool. As a result of the employment of applications using this tool, we have been able to speed up the access of information to the database without having to compromise the detail of the product description field.

¹Standard Industrial Classification Code. A standard coding system used by the federal government to reflect industrial and commercial activities and products of U. S. industry.

²Bureau of Business Research. University of Texas at Austin. Directory of Texas Manufacturers. Vol. 2. Pages 563 and 564.

³Executive Office of the President Office of Management and Budget. Standard Industrial Classification Manual 1987. Pages 211, 212, 214, and 215.

SAS/SC1 software and SAS Macro Facility are trademarks of SAS Institute Inc.