CONSTRUCTING AN INDEX FROM HARDCOPY DOCUMENTS USING SAS

Bob Sturgeon, B.C. Hydro

ABSTRACT

This is a technique for creating an index from a typewritten precis of official "Rate Hearing" proceedings on a diskette. The 1500-page precis consists of precis references and descriptive paragraphs referring to the main hearing documents which are contained in 30 volumes. An index was required for a quick reference to previously discussed items. A three step approach was used. Initially, the precis was optically scanned and recorded on a word processing diskette. After minor editing, the data was transmitted to an IBM paintframe. A SAS was then used to create an index from the dataset.

INTRODUCTION

Rate Hearing applications are an ongoing fact of life for utility companies. A listing was required to provide a quick index to statements or decisions on an issue, made in previous hearings, in order to decrease time and costs involved in manually searching the records to rate hearings for information. A simple method was required to allow searching for terms of a word, and to allow horizontal scanning to remove all special characters if a word contained any numerics. It was to be dropped and line 60 accomplishes this. Line 70 provides a listing of any dropped words (Appendix D). A SAS dataset (SEDATA) was finally created consisting of all words and their page number.

The optical scanner had a habit of randomly inserting binary zeroes into the middle of a word (see Appendix A - Precis h), and Lines 56 to 59 remove these. Lines 51 to 53 allow horizontal scanning to remove all special characters if a word contained any numerics. It was to be dropped and line 60 accomplishes this. Line 70 provides a listing of any dropped words (Appendix D). A SAS dataset (SEDATA) was finally created consisting of all words and their page number.

The exclude file was now read in at lines 74 to 93 and a listing was printed. The index was sorted in lines 95 to 98. A listing was printed containing selected phrases (see Appendix E). This dataset (SEDATA) was finally created consisting of all words and their page number.

The index was reauired in 2 days and SAS was the only language available that could meet this requirement. A facility was provided to search a file of common words that were not required and to delete them from the index. In addition, a listing was required of specified phrases (e.g. cost of living) that might not be on one line but wrap around the end instead.

A facility was provided to search a file of common words that were not required and to delete them from the index. In addition, a listing was required of specified phrases (e.g. cost of living) that might not be on one line but wrap around the end instead.

The Statistical Analysis System was chosen to provide this index for several reasons:

1. SAS lends itself to reading text or lines of data from an ASCII file. A feature would allow selected keywords to be dropped from the index.

2. IBM's 'STAIRS' was available to us but it was not available to another, a hardcopy index that could be carried into the hearing room.

3. The index was required in 2 days and SAS was the only language available that could meet this deadline. PL/I would require 10 to 15 days of programming time, and STAIRS would require 3 weeks to provide an index.

METHOD

1. The typewritten precis (Appendix H) was optically scanned by a CONVUSCAN ALPHAWORD 3 optical scanner.

2. A Word Processing operator edited the data and placed markers (hex '05') to indicate page breaks and 'FIRST' and 'LAST' tests, as well as the 'MISS' feature to indicate page breaks (Appendix A).

3. A dataset of exclusions was created (Appendix B) as a catalog of phrases to be searched for (Appendix C).

4. The SAS program was now used to read the three above-mentioned datasets and create the index. Lines 66-74 of the program read in the precis dataset and perform editing. The SAS 'ARRAY' feature was used to store words in a character string, and the 'MISS' feature was used to set any missing words to blank. Because the data was optically scanned, there was often no indication of the precise location of words. The dataset was then processed to delete any words that did not occur in the original dataset (SEDATA) as defined in Lines 51 to 53.

5. Lines 16-74 of the program read in the precis dataset and perform editing. The SAS 'ARRAY' feature was used to store words in a character string, and the 'MISS' feature was used to set any missing words to blank. Because the data was optically scanned, there was often no indication of the precise location of words. The dataset was then processed to delete any words that did not occur in the original dataset (SEDATA) as defined in Lines 51 to 53.

SUMMARY

The Statistical Analysis System was able to quickly provide an index that would have taken much longer using any other programming language. Several features of the program were only available in the SAS language and would have required extensive programming in other languages, such as the 'MISS' feature, which allowed the selection of phrases that were searched in the preceding step. A listing of any dropped words (Appendix D) was then produced at Lines 162 to 165. A listing (Appendix E) was then produced at Lines 162 to 165. A listing (Appendix E) was then produced at Lines 162 to 165.

ACKNOWLEDGEMENTS

The initial design stage was facilitated by several technical people working on independent projects: Construction of a Literature Index Using SAS was presented at the 1961 SLGCI Conference by Paul J. von Dechren.
SOURCE PROGRAMME

*****************************************************************************
** THIS PROGRAMME WILL READ THE TEXT OF A DOCUMENT THAT CONTAINS A **
** 0 (SPACER) OR A VALID RECORD. DATA WILL BE IN COL 4-60 **
** EXCEPT IF COL 7 CONTAINS A HEX "05", DATA WILL BE IN COL 10-60 **
** IF COL 5 CONTAINS A HEX "05" THEN THE PAGE REFERENCE **
** NUMBER WILL BE FOUND IN COLS 2-4. **
** STANDARD CHARACTERS EXCEPT HYPHEN WILL BE DROPPED AND A NEW DATA SET **
** SET WILL CONTAIN WORDS TO BE EXCLUDED AND THESE WILL BE MATCHED **
** TO THOSE SPECIFIED TO BE EXCLUDED FROM INDEX **
** THIS PROGRAMME WAS ORIGINALLY WRITTEN TO PROVIDE A COMPARABLE READ-TYPED COPY OF A RATE **
** SYSTEM "6" AND TRANSMITTED TO AN ORIG 202-300 DATA SET ON AN IBM MAINFRAME. **
** WRITTEN JUNE 26, 1981 AT BC HYDRO BY BOB STURGEON **
*****************************************************************************

OPTIONS NODATE;
DATA SETNAME;
   KEEP PAGE REF WORD;
   RETAIN PAGE REF;
  长度 = WORD1-WORD70 = 20. ;
   允许从1-70个字符组成20个字符。
   INFILE INEX LS=50 MISSOVER;
   随机读取。
   DATA PHDATA;
   保存特殊字符。
   FILE PRINT LS=72;
   打开文件。
   DATA EXCLUDE;
   INPUT 01 CHECK1 ?? PK1.
   如果 CHECK1 = 5 那么 DO;
   验证页面号。
   END;
   IF CHECK1 = 5 THEN DO;
   十进制。
   END;
   DD OVER WORDS1:
   对于每个单词;
   IF WORDS1 = 0 THEN
   字符串。
   END;
   ELSE
   ORIGINALLY
   DO OVER WORDS1;
   列表。
   END;
   DATA EXCLUDE:
   输入。
   DD OVER WORDS1;
   IF WORDS1 = 0 THEN
   未找到。
   END;
   ELSE
   DATA PHDATA;
   输入。
   DATA EXCLUDE;
   输入。
   PROC SORT DATA=EXCLUDE;
   数据。
   PROC SORT DATA=PHDATA;

864
BY NBRWORDS PHRASE:

DATA EXPAND: * EXPAND TO CARRY PHRASES OF 1 TO 5:
LENGTH PHRASE 270:
KEEP PHRASE NLWORDS:
SET NBRDATA:
NRWORDS = 1:
PAGE = PAGE:
PHRASE = PHRASE:
OUTPUT:
IF NOT END THEN DO:
SET NBRDATA( FIRSTPOS=2) END=END2:
NRWORDS = 2:
PHRASE = TRIM(PHRASE) || TRIM(PHRASE)
IF LENGTH(PHRASE) > 66 THEN
GO TO OUTEX:
OUTPUT:
END:
IF NOT END THEN DO:
SET NBRDATA( FIRSTPOS=3) END=END3:
NRWORDS = 3:
PHRASE = TRIM(PHRASE) || TRIM(PHRASE)
IF LENGTH(PHRASE) > 66 THEN
GO TO OUTEX:
OUTPUT:
END:
IF NOT END THEN DO:
SET NBRDATA( FIRSTPOS=4) END=END4:
NRWORDS = 4:
PHRASE = TRIM(PHRASE) || TRIM(PHRASE)
IF LENGTH(PHRASE) > 66 THEN
GO TO OUTEX:
OUTPUT:
END:
IF NOT END THEN DO:
SET NBRDATA( FIRSTPOS=5) END=END5:
NRWORDS = 5:
PHRASE = TRIM(PHRASE) || TRIM(PHRASE)
IF LENGTH(PHRASE) > 66 THEN
GO TO OUTEX:
OUTPUT:
END:
OUTEX:
PROC SORT DATA=EXPAND;
BY NRWORDS PHRASE PAGE:
DATA EXPAND:
SET EXPAND:
IF FIRST PAGE;
DATA EXPHRASE:
READ PHRASE(IN=PH) EXPAND(IN=EX);
BY PHRASE:
IF PH & EX;
* WAS PHRASE FOUND:
PROC SORT DATA=EXPHRASE:
BY PHRASE PAGE:
DATA _NULL:_
SET PHRASE:
* PRINT PHRASE REFERENCES:
FILE PRINT LS=72 LINESLEFT=L COLUMN=C;
SET PHRASE:
* RUN ON "GP CLASS WITH CHAR=GB10:
GET TOPLN: 50:
* SET UP FIRST COLUMN VARIABLE POSITION:
OUT REFRENCE INDEX OF SELECTED PHRASES:
TITLE2 B.C. UTILITIES COMMISSION HEARING:
TITLE3 PUBLIC CONVENIENCE AND Necessity:
TITLE4 SMALL RIVER HYDROELECTRIC REDEVELOPMENT PROJECT:
TITLE5 HYDRO DEVELOPMENT PROJECT:
TITLE6 "RIVER HYDROELECTRIC REDEVELOPMENT PROJECT:
TITLE7 FEDERAL APPLICATION FOR A CERTIFICATE OF:
TITLE8 FEDERAL APPLICATION FOR A CERTIFICATE OF:
TITLE9 FIRST PHRASE THEN DO;
* START OF NEW PHRASE:
IF L < 5 THEN
PUT "PHRASE":
PUT "PHRASE"
PUT PAGE:
PUT PAGE:
PUT "PHRASE" 2:
PUT PAGE:
GO TO LAST:
PUT "NEW PAGE":
* NOT FIRST PHRASE, JUST PRINT PAGE:
PUT PAGE:
IF FIRST.PHRASE THEN DO;
* LAST PAGE REF SO FREE THE PRINT LINE:
POT RETURN:
END:
PUT DC="1"
IF L > 6 THEN DO;
* NOT LAST PAGE REF SO MOVE ON "":
END:
PUT DC="1"
IF L > 6 THEN DO;
* NOT LAST PAGE REF SO MOVE ON "":
END:
RETURN;
END:
PROC SORT DATA=SETDATA:
BY WORD PAGE:
DATA SETDATA:
SET SETDATA:
IF FIRST.PAGE:
DATA NEW:
READ SETDATA(IN=N) EXCLUDE(IN=X):
BY WORD;
865
APPENDIX "A"

B.C. UTILITIES COMMISSION HEARING
B.C. HYDRO 1981 APPLICATION FOR A CERTIFICATE OF
PUBLIC CONVENIENCE AND NECESSITY
SMALL RIVER HYDROELECTRIC REDEVELOPMENT PROJECT
5 May 1981

Pace Ref.

1 Appearance of counsel for B.C. Hydro and the Department of Fisheries and Oceans were listed.

2 Mr. Johnston for B.C. Hydro gave opening remarks on the presentation of the application. He identified the presence of witnesses, and requested the opportunity to have support personnel sit with or behind the witnesses to provide technical or detailed material. Johnston proceeded to call his first witness when Mr. Smith, the Commission counsel, interrupted to determine the status of the supporting documents.

3 Johnston said since the documents were referred to in Mr. Jones's testimony, he did not propose to file them as evidence. Smith was concerned the single application unlikely contained sufficient information for the Commission to reach a conclusion. Johnston stated the form of the application had caused difficulty and he would file the documents if requested by the Commission. Following some discussion, Johnston was instructed the required number of copies was available and he suggested they be filed.

APPENDIX "B"
APPENDIX "H"
B.C. UTILITIES COMMISSION HEARING
B.C. HYDRO 1981 APPLICATION FOR A CERTIFICATE OF
PUBLIC CONVENIENCE AND NECESSITY

SMALL RIVER HYDROELECTRIC REDEVELOPMENT PROJECT

5 May 1981

Page Ref.

1.Appearances of counsel for B.C. Hydro and the Department of Fisheries and Oceans were listed.

2. Mr. Johnston for B.C. Hydro gave opening remarks on the presentation of the application. He identified the sequence of witnesses and requested the opportunity to have support personnel sit with or behind the witnesses to provide technical or detailed material. Johnston proposed to call his first witness when Mr. Smith, the Commission counsel, interrupted to determine the status of the supporting documents.

5. Johnston said since the documents were referred to in Mr. Jones's testimony, he did not propose to file them as evidence. Smith was concerned the single application unlikely contained sufficient information for the Commission to reach a conclusion. Johnston stated the form of the statute had caused difficulty and he would file the documents if requested by the Commission. Following some discussion, Johnston was instructed the required number of copies was available and he suggested they be filed.