SMASER

# EXIF

TRIPsystem
Product Documentation

# End User License Agreement

All rights to this software, its documentation and logotypes of the TRIP product family and software (altogether “Software”) supplied by Smaser AG (Smaser) are exclusively owned by Smaser.

The transfer of this Software, solutions or parts thereof requires the prior written agreement of Smaser. Furthermore, the customer has the right to use licensed Software and / or process solutions supplied by Smaser to the extent specified in his contract with Smaser.

The free-to-use non-commercial version doesn’t require a prior written agreement with Smaser but such customers, organizations and/or third parties agree by using the software and / or solution of Smaser to be strongly obliged to keep all rights to this software, documentation and logotypes of the TRIP product family absolutely uninfringed and protected.

## About this document

The following document describes the prompts that you will get from Exif, and how to respond to them to get differing kinds of output. The purpose of the exercise is to help a customer support representative to be more effective in the isolation of index related problems prior to bug reporting.

## Conventions used in this document

| | |
|---|---|
| **Xyzzy (x)** | is a prompt from EXIF to the user (where 'x' is the default response provided by EXIF and taken if the user simply hits the RETURN key) |
| ***If <clause> ... Else ...*** | depending on the response given to one or more prompts, certain further prompts will either be given or not. The conditional clause specified after the word "If" determines the appearance or otherwise of those prompts. If the conditional clause is not matched, those prompts shown after the "Else" word are given. |
| <tab> | i.e. an indent is used to give logical bounding to prompts where a group of prompts will only appear if the pre-stated condition is true. |

## What is EXIF?

EXIF is a very powerful utility for the examination of TRIP index files. This utility can be used to locate and detect almost any bug which has any relevance to the indexing, searching and highlighting functions in the TRIP system.

With practice it becomes simple to understand what EXIF is reporting, and how that differs from what should be reported, as defined by your knowledge of what is, or should be, in the database itself.

The EXIF utility program is located in the directory with the same path that is defined as TDBS_EXE in the TRIPsystem configuration file tdbs.conf. Run EXIF from the command line:

```
<TDBS_EXE>/exif
```

## Command line options

EXIF takes the following options on the command line (see under "Prompts from EXIF" for more detailed explanation):

| | |
|---|---|
| -f <filename> | BIF / VIF Filename |
| -t <typenr> | Low type number (2-15) |
| -T <typenr> | High type number (2-15) |
| -n <level> | Level number (1-6) |
| -l <value> | Low level value (0-1000000) |
| -L <value> | High level value (0-1000000) |
| -i <value> | Low initial number (33-255) |
| -I <value> | High initial number (33-255) |
| -G <values> | Comma-separated list of GBK character byte pairs |
| -N <typenr> | Type number (if --one-term is specified) |
| -v <termvalue> | Term (if --one-term is specified) |
| -F <fieldnr> | Field number (if --no-one-term is specified) |
| -r <rid> | First record number (if --no-all-occs is specified) |
| -R <rid> | Last record number (if --no-all-occs is specified) |
| -e <blocknr> | First entry block number |
| -E <blocknr> | Last entry block number |
| -O <limit> | Occurrence count limit (1-10000) |
| --[no-]tstamp | Enable or disable use of timestamp limit |
| --[no-]normal | Normal run? |
| --[no-]one-term | One term only? |
| --[no-]select | Select terms? |
| --[no-]lvl-restrict | Level restriction? |
| --[no-]gb2312 | Show GB-2312-80 codes? |
| --[no-]all-occs | All term occurrences? |
| --[no-]segsum | Segment summaries only? |
| --[no-]stg-stat | Storage statistics? |
| --[no-]sum-stat | Summary statistics? |
| --[no-]min-sum | Minimal summary? |
| --[no-]one-entry | One entry block? |
| --[no-]entry-sum | Entry block summaries? |
| --[no-]term-stat | Statistics for each term? |

Note that option names are case sensitive. In addition EXIF takes the following option that assigns reasonable defaults to options whenever possible:

| | |
|---|---|
| --defaults | Use defaults for options when possible |

Options specified on the command line will not be prompted for unless the associated specified value is invalid.

## Prompts from EXIF

**BIF / VIF Filename :**

Specify here the actual filename for the index file that you wish to examine. You cannot specify a database name, as this utility expects to be used when, for example, CONTROL is corrupted.

**Time stamp limit (N) :**

Allows you to specify a time before which entries in the index file will not be considered a part of this EXIF run. This can be useful when, for instance, a customer can give a date and time of the last successful index. Obviously, only updates from this time onwards should be examined when looking for index errors.

***If YES to "Time stamp limit"***

**Date & time (e.g. 890423 234225) :**

Allows you to specify the actual date and time (in the format shown in the example) at which you want EXIF to begin its examination. For instance, in the example shown, only updates to the index occurring after 11:42:25 p.m. on the 23rd of April 1989 would be considered relevant to this run.

**Normal run? (Y) :**

The definition of a "normal" run is rather fuzzy, but basically it bypasses many of the prompts and selects default values for certain run parameters. Effectively, it :

- Disallows selection of specific terms
- Disallows the facility to examine the posting list for terms
- Disallows the facility to gather extended storage statistics for terms

***If NO to "Normal run"***

**One term only? (N) :**

Allows you to specify that you wish to examine the extended attributes of a particular term in the index file. If you do not wish to examine extended attributes, you would normally have chosen to use a "Normal run" (see prompt above). Your response to this prompt would thus normally be YES.

***If YES to "One term only"***

**GB-2312-80 codes to be shown? (N) :**

Allows you to search for Chinese / Japanese pictograms by decimal value. In fact, it allows you to search for any term in the index which has a two byte length and whose lead byte is greater than 128. Normally, however, this only has meaning when searching for pictograms. Your response to this prompt would thus normally be NO.

**Type number (2) :**

Allows you to specify the type of term for which you want to search (and subsequently have displayed). These types are roughly similar to normal TRIP field types, but there are some important differences. The currently supported term types are shown in the table below.

| Type | File Type | Term(s) to be searched / displayed |
|---|---|---|
| 2 | BIF | word value |
| 6 | VIF | word value in phrase value |
| 7 | VIF | n-gram value |
| 9 | BIF | NUmber field |
| 10 | BIF | INteger field |

| | | |
|---|---|---|
| 11 | BIF | DAte field |
| 12 | BIF | TIme field |
| 13 | BIF | complete phrase value |

There are several things to note here.

Firstly, type 2 covers both TExt and PHrase field types. This is the term type used when performing general searches in TRIP (Find xyz).

Secondly, there is a great difference between types 9,10,11,12 and all others. Note that the description of these types says “xxx field” rather than “xxx value”. This is because in the BIF the indexed entity for a numeric field type is that field’s number, not the values that it contains.

Hence it is not possible to use EXIF to locate a particular numeric value without performing significant post-processing (as described below in the section concerning “Field number”).

Finally note that this points out just how “fatly” PHrase fields are stored. First they are stored as data in the BAF, then as a complete phrase in the BIF (type 13), then each word from the phrase is stored as a word in the BIF (type 2), then each word is stored as a word in the VIF (type 6), finally each n-gram from each word is stored in the VIF (type 7). This rather significant overhead provides wonderful search capability, but does highlight the fact that use of PHrase fields must be approached cautiously.

***If NO to “GB-2312-80 codes” )***

***If RESPONSE to “Type number” was 2, 6, 7 or 13***

**Term :**

Allows you to type in a literal term to be searched for using the type number supplied earlier. Note that in the case of type 7, the term must either be a single character, a pair of characters or a triplet of characters. In all cases, the term is normalized before being searched for.

***Else ( RESPONSE to “Type number” was NOT 2, 6, 7 or 13 )***

**Field number :**

In the case of numeric fields, types 9 - 12, the indexed entity in the BIF is the field number, not the field content. Therefore, EXIF will prompt for the field number as reported by the CCL command STatus.

In order to determine whether a particular field value has been indexed, you must request a complete output of the field’s content using the prompt “Segment summaries only?” as described below. Then, you must manually search the list of output values to determine whether the value in question has been correctly added to the BIF.

***Else (YES to “GB-2312-80 codes” AND any RESPONSE to “Type Number”)***

**First byte (256-done) :**
**Second byte :**

These two prompts allow you to specify the decimal value of the particular pictogram for which you are searching. Note the prompts repeat until you specify 256 for the “first byte” but at present all

pictograms are stored as single two byte entities and therefore the second response to “first byte” should always be 256.

This prompt can be useful when examining a database created in a different character set than that which you are using to run EXIF. For example, if the database was created using ISO Latin 2 (South Eastern Europe), and you are examining it using DEC Multinational, you should use these prompts to specify the decimal byte values of the term in question, rather than typing in the term literally. Obviously, this only works properly when examining terms which contain an even number of characters.

**All occurrences? (Y) :**

Allows you to specify whether EXIF should process all occurrences of the specified term, or whether you wish to supply a particular record range to examine.

***If NO to “All occurrences”***

**First record number :**
**Last record number :**

These two prompts allow you to define the range of records within which you wish to examine a particular term’s postings.

**Segment summaries only? (Y) :**

Allows you to specify whether you want to see the content of the individual postings for the term, or whether you simply wish to get an overview of the segment layout in the index file.

***Else (NO to “One term only”; Internal R&D use only, always specify YES to “One term only”)***

**Select terms (Y) :**
(plus a number of other questions...)

***Else ( YES to “Normal run” )***

**Storage statistics? (Y) :**

Allows you to specify whether you want to see a breakdown of how much space is being occupied by terms in the index. If you specify NO to this prompt, you will simply see a list of terms from the entry blocks specified below. If you specify YES, you will get to see how the index entries are laid out in terms of bytes used, levels of data segments and pointers, etc. This can be useful when used in conjunction with the REBIF utility described in document TRIP-REBIF.

**Summary statistics? (Y) :**

Allows you to specify that you wish to see a summarized version of the storage statistics.

***If YES to “Summary statistics”***

**Minimal summary? (Y) :**

Allows you to request the smallest possible amount of information from EXIF - very little use normally.

**One entry block? (Y) :**

Allows you to specify that you wish to examine a particular block in the primary index table. This is of more use to R&D than to agents, although it can be interesting to see

which terms collide with each other in the index itself.

***If YES to "One entry block"***

**Please enter one of its terms**

Type in a term. The output from EXIF will now show not only that term but also any other terms in the database which "collide" with that term in the index. There will be many of these.

***Else ( NO to "One entry block" )***

**First entry block number (2) :**
**Last entry block number (max. entry block in index) :**

Allows you to specify a range of entry blocks in the index table to examine. If you wish to examine only one block, and do not know any of its terms, use this route rather than the one shown above, and simply specify first and last as the same number.

In order to examine the entire index table, accept the default values for both prompts. In this case, if you have responded YES to "Minimal summary", you will also see output showing the layout of so-called Free Classes within the entry table. This output is useful when optimizing the performance of a database, particularly for CD-ROM use, and is described more in the documentation for the REBIF utility.

***If first entry block <> last entry block AND NO to "Minimal summary"***

**Entry block summaries? (Y) :**

Allows you to specify that you want summary output at the end of each entry block in the display.

***If NO to "Summary statistics"***

**Statistics for each term? (N) :**

Allows you to specify that you want to examine the terms themselves, i.e. to get a list of what is in the index along with storage information (if you asked for it when prompted "Storage statistics").

***If NO to "Statistics for each term"***

**Occurrence count limit :**

Allows you to set a limit above which terms will be displayed in full and below which terms will simply be summarized.

## "Real world" examples of using EXIF

You want to see whether a term has been indexed or not. You might not be able to find it with TRIP and you want to know whether it simply isn't there, or whether the index might be corrupt. This can be ascertained with Exif because its entry block search mechanism is much less optimized than is that used by the normal TRIPkernel search routines.

```
*** TRIP System Utility EXIF - BIF/VIF examination ***
     Version 6.0-0   10-Dec-2008 16:21:48.62

BIF or VIF file name                 : trip$demo:alice.bif

Time-stamp limit?                (N) : n
Normal run?                      (Y) :
```

```
Storage statistics?               (Y) :
Summary statistics?               (Y) : n
One entry block?                  (Y) : y
Enter one of its terms                : alice
Statistics for each term?         (N) : y

Term            Type   Ocnt  Rcnt/F    Sspace  Pcnt    L0    L1    L2   L>2
--------------   -- ------- ------- --------- ----- ----- ----- ----- -----

...

ALICE             2     923     426      2331     2     -    19     -     -

...
```

In this extract from the Exif output, we can easily see that the term in question has been indexed. It occurs 923 times in 426 records and takes up 2331 bytes. That storage space is divided between two down level pointers, and 19 level 1 posting segments.

Following from the previous example, we now want to know whether a particular occurrence of the term is indexed. Let us say that a search for "alice" is resulting in many hits, but you can see from the Show command that one, or more, of the terms is not getting highlighted and you now want to determine whether this is because the term has not been indexed, or because TRIP has a bug in its output formatter. First, therefore, we look at the segments (19 of them from example 1) to see if the record in question actually is recorded in the BIF :

```
*** TRIP System Utility EXIF - BIF/VIF examination ***
     Version 6.0-0   10-Dec-2008 16:29:11.14

BIF or VIF file name                  : trip$demo:alice.bif

Time-stamp limit?                 (N) :
Normal run?                       (Y) : n
One term only?                    (N) : y
GB-2312-80 codes to be shown?     (N) : n
Type number:                      (2) : 2
Term:                                 : alice
All occurrences?                  (Y) :
Segment summaries only?           (Y) :

Pointer (1) 788 occurrences, 376 records in ofl. block 20; Rn=1-420
  ...
Pointer (1) 135 occurrences, 50 records in ofl. block 21; Rn=421-473
  Segment  17:   21, 49 in 117 bytes,   Rn=421-443
  ...
```

This lets us see that if, for instance, our non-showing term was in record 438, this record is covered by a data segment in level 1 (segment 17). This segment contains 21 records, 49 occs in 117 bytes. Segments marked with an asterisk (not actually shown here) simply show that a particular record's occurrences bound more than one segment.

Again, following from the previous example, we now want to narrow our search down so that we can see the actual postings for the record in question - 438. We could do this by asking for all occurrences, and then scanning the list - or we can simply specify the exact record that we are interested in :

```
*** TRIP System Utility EXIF - BIF/VIF examination ***
     Version 6.0-0   10-Dec-2008 16:34:43.97

BIF or VIF file name                  : trip$demo:alice.bif

Time-stamp limit?                 (N) :
Normal run?                       (Y) : n
One term only?                    (N) : y
GB-2312-80 codes to be shown?     (N) : n
Type number:                      (2) : 2
Term:                                 : alice
All occurrences?                  (Y) : n
First record number?              (0) : 438
```

```
Last record number?              (438) :
Segment summaries only?            (Y) : n

    438      0      2      1      1      2
    438      0      5      3      1     17
    438      0      5      5      1     14
    438      0      5      7      1      8
```

The output from this run needs some explanation. Each field type yields different meaning output in the following manner :

Type 2 (word in BIF)
<record #> <part #> <field #> <paragraph #> <sentence #> <word #>

Type 6 (word in PHrase within VIF)
<internal identifier> <not used> <word # in phrase> <not used>

Type 7 (n-gram within VIF)
<internal identifier> <not used> <gram offset in word> <not used>

Type 9, 10, 11, 12 (NUmber, INteger, DAte, TIme in BIF)
<record #> <part #> <field #> <subfield #> <value> <not used>

Type 13 (PHrase in BIF)
<record #> <part #> <field #> <not used> <subfield #> <not used>

Thus we can see that the term "alice" occurs in the record 438 a total of four times :

1. Part 0, field 2 (CHAPTER), para 1, sent 1, word 2
2. Part 0, field 5 (TXT), para 3, sent 1, word 17
3. Part 0, field 5 (TXT), para 5, sent 1, word 14
4. Part 0, field 5 (TXT), para 7, sent 1, word 8

Note that in the case of the first occurrence, in the field CHAPTER, the value for the paragraph number is ignored and the value for the sentence number is used as the subfield number.

Note also that part 0 indicates the head record.

Now that you examine a particular term's occurrences to this detail, you can determine very quickly where a suspected problem lies - either with the indexing system or with the output formatter or with the search system.

Another request from customers tends to be to find out how much storage overhead is being incurred for indexing common terms such as "THE". There is no direct function to return this, but again Exif produces the goods!

```
*** TRIP System Utility EXIF - BIF/VIF examination ***
     Version 6.0-0   10-Dec-2008 16:49:44.93

BIF or VIF file name                 : trip$demo:alice.bif

Time-stamp limit?                (N) :
Normal run?                      (Y) : n
One term only?                   (N) : y
GB-2312-80 codes to be shown?    (N) : n
Type number:                     (2) : 2
Term:                                : the
All occurrences?                 (Y) :
Segment summaries only?          (Y) :

... lots of segment summaries appear ...

THE              2    70 segments    3223 occurrences    7696 bytes
```

The last line of output from Exif is a complete summary of the storage statistics for the term in question, in this case "THE". We can see here that to store the type 2 version of "the", it takes 70 data segments to hold the 3223 occurrences of the term, which has been compressed to 7696 bytes (3223 type 2 occurrences uncompressed would take up 77352 bytes, a compression rate of approximately 91%).

To get a complete picture of how much space is being used by a term, you should also look for it as type 13 (complete PHrase), and as type 6 in the VIF (word in PHrase). Simply add all of these numbers together to get a final report of the damage. Note that looking for "THE" as type 7 in VIF should not be contributed to the total as this would be including the storage space for the n-gram within terms such as "THERE", "THEIR", etc.

This can also be produced by using the following sequence :

```
*** TRIP System Utility EXIF - BIF/VIF examination ***
     Version 6.0-0   10-Dec-2008 17:07:49.86

BIF or VIF file name                  : trip$demo:alice.bif

Time-stamp limit?                (N) :
Normal run?                      (Y) :
Storage statistics?              (Y) :
Summary statistics?              (Y) : n
One entry block?                 (Y) :
Enter one of its terms               : the
Statistics for each term?        (N) : y
```

which will yield output for the entire entry block containing the term in question. The advantage of doing this is that both BIF types of the term will be listed underneath each other along with the total storage space for that type (column Sspace).