
MIR TUTORIAL ONE
DATABASE ANALYSIS
Detailed Table of Contents

0. MIR TUTORIAL ONE Table of Contents

1. Introduction to MIR TUTORIAL ONE

1.1 Project overview

1.2 Tutorial ONE overview

2. Source code guidelines

2.1 Needs of the information searcher

The value of time
Simplicity, simplicity, simplicity
Control
Freedom from a ticking clock
Freedom from obscure error messages
Freedom from the curse of codes
Language of choice
Context-sensitive help
More bang per computer dollar

2.2 Design background

Squeezing each bit... the conservationist start
The gigabyte years
Unix influence
C with a Fortran accent

2.3 Design decisions

Language
Hardware
Operating system and compiler
Avoiding code that blows up

2.4 Conventions

Humans use programs
Humans read programs

2.5 Use It, Improve It

3. Data gathering

3.1 Some definitions

Datum
Data
Record
Information
Knowledge

3.2 Why gather data?

3.3 Who are data gatherers?

3.4 Keyboard data input

3.5 Scanned data input

3.6 Formats, standards and common sense

3.7 Data quality

Accuracy
Timeliness
Consistency

3.8 Value of data

PRE-RELEASE MAY 13 92

Tables of Contents may be copied provided no changes are made.
MIR TUTORIAL ONE Copyright (C) Marpex Inc., 1992 Page 0.

- Market capacity
- Cost recovery strategy
- Educating the market
- Perception of value
- Value added through combination

3.9 Data ownership

3.10 Summary

4. First steps in data analysis

4.1 Objectives

- Extract searchable content
- Recognize record separations
- Recognize field separations
- Recognize formatting aids

4.2 Learn how the data was accumulated

4.3 Learn how the data will be used

4.4 Access to samples and hard copy

- Media
- Representativeness
- Hard copy

4.5 Access to software tools

4.6 Extracting samples from larger files

- Use CPB to get subsets

4.7 Byte surveys - a worked example

- A_BYTES to analyze bytes

Sorting byte analysis reports
A_BYTES -L for locations data

4.8 Data types

- ASCII text
- Extended ASCII text
- Text with ASCII markup codes
- Text with binary markup codes
- Text with packed numbers
- Text with compression substitutions
- EBCDIC
 - EBC_ASC to convert EBCDIC to ASCII
- Binary data

4.9 Data presentation

- Byte stream
- Line records
- Fixed length records
- Blocked records with ASCII lengths
- Blocked records with binary lengths

4.10 Byte distributions

- English text
- European languages text
- Significance of byte frequencies

5. Patterns in byte sequences

5.1 Heads and tails... first impressions of a file

- HEAD to see the beginning of a file
- HEAD ## to see ## lines
- HEAD -t to see the tail end of a file
- HEAD -a to see accented characters

5.2 Non-DOS files

- DOSIFY to insert carriage returns

5.3 Displaying printable data

F_PRINT filter

5.4 Detailed data dumps

DUMP to display hex and ASCII

5.5 Convenient display of fragments

FRAGMENT to show context

5.6 Viewing patterns throughout a file

A_PATRN to extract byte patterns

5.7 The power of sorting patterns

5.8 Sorting large files

SORT2 for files over 60k

COLRM to reduce large files before sorting

A_OCCUR to analyze occurrences

A_OCCUR2 and A_OCCUR3 utilities

6. Worked Examples - Variations in ASCII text

6.1 Other analysis tools

LINES for a quick line count

A_LEN for a distribution of line lengths

LINE_NUM to insert line numbers

6.2 ASCII markup patterns

6.3 Standard Generalized Markup Language (SGML)

6.4 Free versus hierarchical text

PRE-RELEASE MAY 13 92

Tables of Contents may be copied provided no changes are made.

MIR TUTORIAL ONE Copyright (C) Marpex Inc., 1992

Page 0.

6.5 Fielded variable length text

6.6 Independent versus continuous data

7. Worked Examples - Fixed length records

7.1 Recognizing fixed length ASCII text

NEWLINES to separate records

7.2 Field layouts

7.3 Extracting a single field

7.4 Packed numbers in fixed length records

8. Worked Examples - Binary data

8.1 The preprocessing option

8.2 File signatures

8.3 Converting word processing files

8.4 Binary deblocking lengths

HEX_BIN to create test files

8.5 Binary data in fixed length records

8.6 Compressed data

9. Data Deblocking

9.1 An aid in analysis

9.2 Reducing line records

F_TRAIL

9.3 Handling fixed length records

P_FIXED

9.4 Blocked records with ASCII lengths

DEBLOC_A

9.5 Blocked records with binary lengths

DEBLOC_B
P_MARC

10. Glossary and index of terms

END OF MIR TUTORIAL 1

PRE-RELEASE MAY 13 92

Tables of Contents may be copied provided no changes are made.

MIR TUTORIAL ONE Copyright (C) Marpex Inc., 1992

Page 0.