++
++++

4. HOW THE MIR PROJECT WORKS FOR YOU

++
++++

++++++++++++++++++++++++++++
4.1 "Free" software
++++++++++++++++++++++++++++

In the MIR project we are using the "copyleft" strategy of the Free Software Foundation. The Foundation's GNU General Public License is included as Topic Five; it applies to all software created as part of the MIR project. This software has been created specifically for this purpose by Marpex Inc. since March 1991.

The Free Software Foundation

> "is dedicated to eliminating restrictions on copying, redistribution, understanding, and modification of computer programs. [They] do this by promoting the development and use of free software in all areas of computer use... 'Free' pertains to freedom, not to price... You have two specific freedoms once you have the software: first, the freedom to copy the program and give it away to your friends and co-workers; and second, the freedom to change the program as you wish, by having full access to source code. Furthermore, you can study the source and learn how such programs are written. You may then be able to port it, improve it, and share your changes with others."

What is copyleft?

> "The simplest way to make a program free is to put it in the public domain, uncopyrighted. But this allows anyone

> to copyright and restrict its use against the author's wishes, thus denying others the right to access and freely redistribute it. This completely perverts the original intent.
>
> "To prevent this, we copyright our software in a novel manner. Typical software companies use copyrights to take away your freedoms. We use the copyleft to preserve them. It is a legal instrument that requires those who pass on the program to include the rights to further redistribute it, and to see and change the code; the code and rights become legally inseparable."

Quotes in the above three paragraphs are from page 3 of the January 1992 "GNU's Bulletin" semi-annual newsletter of the Free Software Foundation, 675 Mass Avenue, Cambridge, MA 02139 USA.

The argument for this strategy is set out nicely in an article "Programs to the People" in the February/March 1991 issue of the M.I.T. Technology Review. With permission of the author, Simson L. Garfinkel, the text of the article is included in a separate file on the CD-ROM release(s). The file is named "TOPEOPLE".

+++++++++++++++++++++++++++++++++++

4.2 Interactive publishing

+++++++++++++++++++++++++++++++++++

These tutorials are part of the ongoing output of the MIR project. MIR is an acronym for Mass Indexing and Retrieval. MIR seeks to raise the quality of information search in large masses of data. Seed funding was provided by the Canadian government with the understanding that the underlying indexing and retrieval techniques developed in the project shall be made broadly available under copyleft rules. Personnel from two companies are carrying out the project.

Innotech Inc. of Scarborough, Ontario (416 321-3838) aims toward excellence in CD-ROM publishing services. It is developing interfaces and applications based on MIR technology. Innotech offers consulting services as well as service bureau processing in CD-ROM publishing.

Marpex Inc. is a firm founded in 1976 by the author of the tutorials and the related software. Marpex developed the techniques and pilot programs for the pioneering FindIT CD-ROM system, and more recently collaborated in the design of the Discis Knowledge Research CD-ROM books. Marpex provides consulting in records management, and seminars related to the techniques in the MIR tutorials.

MIR tutorials are designed to be an exercise in co-operative development. The tutorials are being released in five parts. The purchase of any tutorial entitles the buyer to the source code and DOS executable version of the software related to the tutorial. We hope to engage you, the readers and users, in the project. We know that co-operative development will lead to improved end results; many minds are better than one. Text and software is modified according to your input... clarifications, improved methods, more powerful source code, etc. Each tutorial will evolve to reflect significant improvements, with your name attached to the improvements you provide.

After the interactive phase is over, we plan to compile a reference text based on the tutorials. This will be accompanied by a CD-ROM containing all software and support files. Since ISO 9660 CD-ROMs are operating system independent, your ported versions of programs can be included.

Why not release everything at once? Reasons for progressive releases are:

> Scope of the project: Look at the table of contents. There is simply too much for one tutor to complete in a single step. Extensive new work is continuing to be carried out; we are not carrying forward a single line of source code from any proprietary system. Much of this work in the past has been on UNIX workstations; now we are achieving levels of efficiency that can make preparation of large databases feasible on a personal computer.

> Market readiness: Until the first two tutorials have been on the market for a few months, we do not know if our target groups are sufficiently interested. We want to know that our work is meeting a genuine need and that co-operative development under "copyleft" rules is viable.

> Financing: The Canadian government provided seed funding, that is, enough to get the project off to a good start. We are using the same

approach as the Free Software Foundation to provide the money required to carry the project forward. Their major financing is through distribution of tapes containing their work - at roughly $200 for each of several tapes. We aim to carry forward the MIR project through selling paper copies of the tutorials at a very attractive price... $95 for the first copy, $49 per additional copy in the same shipment. Each purchase is accompanied by a free copy of the latest version of the related software. People are free to make copies of the source code and executable programs. We trust that buyers will honor the copyright of the tutorials. If you need extra copies, please buy them from us. We appreciate your support, enthusiasm and encouragement!

++

4.3 Engine-independent techniques

++

The ISO 9660 CD-ROM standard and Microsoft's MS-DOS extensions opened the way to accessing the files on any conforming CD-ROM. But having access to files is not the same as being able to search conveniently. Because indexing systems and <u>interfaces</u> are proprietary, the user has been faced with the nightmare of having to learn a new retrieval method every time a CD-ROM title is purchased from a new vendor. The plea goes up: "Why can't I use the same program I've already learned?"

Why not, indeed?

Two ideas have emerged in the literature. One is full "<u>interoperability</u>"... the ability for a person to select her/his own preferred retrieval interface software and use it to search within any CD-ROM title on any CD-ROM drive under any operating system.   That's far off yet. The second idea, a subset of the first, is now before a Standards Committee (SCAD) of the International Standards Organization (ISO) and may show up in commercial products in 1993. That is the possibility of <u>separating the software into a client interface and an underlying server</u> which fetches data from the CD-ROM. The server module resides in RAM and communicates with the client interface through standardized ASCII strings. The intention is that the server is specific to the data and the indexes in place; the client interface is the user's preference of any retrieval software conforming to the standard.

These engine-independent techniques do away with the high cost and inconvenience of re-education. There are perhaps five contending proposed standards. The Information Handling Committee of the Intelligence Community Staff in Washington, D.C. has commissioned the CD-ROM Read-Only Data Exchange Standard (CD-RDx). The aircraft industry appears seriously committed to Structured Full-Text Query Language (SFQL), an extension of the ISO approved SQL. Other contenders are V39.50 (a library system networking protocol), Silver Platter's DXS, and DFL, an earlier outgrowth of Standardized Query Language. Unknowns at this point include the data structures supported (whether columnar relational databases and subsets thereof, or whether more generalized forms), and the actual syntax of messages that pass between the interface and server modules.

We believe that cooperative development through the MIR project can contribute to this process. If software is freely available under copyleft rules, it can be adapted very readily as standards evolve. No-one has to hold back until the Standards Committee makes its one year or three year or five year report.

We also believe that it is unnecessary to limit the discussion to CD-ROM. The basic problem (frustration at being forced to learn new interfaces) is independent of the medium on which the data are stored. MIR technology may be applied to data held on hard disk, floppy diskettes, Write Once Read Many (WORM), Bernoulli, rewritable laser optical disks, laser cards or whatever other media can retain data as byte streams.

++++++++++++++++++++++++++++++++++

4.4 The software provided

++++++++++++++++++++++++++++++++++

Scope: The source code for data analysis and preparation, search term selection, and to some extent automated indexing require little interaction with a user. The programs in TUTORIALS ONE through THREE are therefore considered complete.

TUTORIAL FOUR presents an engine (a "data server module") which may be used with interfaces compatible with engine-independent techniques. The number of different interfaces that might be written is infinite. Interface source code can be (and is likely to be) handled in traditional proprietary ways, simply because of the great variability in features that end users desire. You or your firm may write a "client module" interface and keep it proprietary, provided the data server module is kept separate and under copyleft rules. If you care to write a client module under copyleft rules, and if it works well, we will be glad to pass it along.

The software provided with TUTORIAL FIVE might be classed as "discussion starters". We carry the discussion a fair distance, but look to readers to pursue their specific interests. In an ideal world, that pursuit would take the form of a public exchange of ideas under copyleft rules. As Captain Jean-Luc Picard would say, "Make it so!"

Naming conventions are applied to many of the programs. DOS constrains source code names to eight characters plus a ".C" extension. Where a six letter name is workable, a single letter followed by an underscore precedes the name and has one of the following meanings:

A_*.C	analyze, report
B_*.C	build indexes
C_*.C	compress / integerize data
E_*.C	expand content of a file
F_*.C	filter out parts of a file
I_*.C	invert token matrix
J_*.C	join words into useful phrases
M_*.C	merge files
P_*.C	pre-process particular layouts
Q_*.C	quality assurance
R_*.C	rotate content within a line
S_*.C	server module for retrieval
T_*.C	transliterate language to ASCII

Support files include LICENSE.WP and LICENSE.ASC. These WordPerfect 5.1 and ASCII versions of the Free Software Foundation's GNU General Public License govern permissions for software supplied with the tutorials. You will find an ORDER form, again in WordPerfect and ASCII versions. CD-ROM release(s) contain extra worked examples, and articles such as TOPEOPLE.

We recommend you place executable copies of all programs in one area on your hard disk. That way, you can create easy access to the programs with only one small addition to your DOS path (something of the form "\C:\BIN;" added to the PATH line in your AUTOEXEC.BAT file).