++
++++

3. WHY YOU WANT TO USE MIR TUTORIALS

++
++++

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

3.1 The end user and the thirst for knowledge

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

People need to know. To know is to have understanding. To know is to recognize the nature of something going on in our world. For each of us, knowing is the key to control over our environment. To know is to gain self-esteem and confidence. Knowledge equips a person to create value. And creating or adding value is what our working life is about. Knowledge is always the first step.

The objective of computerized indexing and retrieval is to serve people's need for knowledge. The objective is NOT tidy techniques; it is service and empowerment. Efficient techniques are simply a means to the end. For those who insist that we focus on profit and the bottom line, consider this: If we keep improving in our recognition of human need and our service of that need (and if we don't "park our brains at the door" in the process), that is the surest way to ongoing profit.

Everything that follows seeks to give control to the end user. Knowledge itself increases a person's control over his or her world. The tools that we put in the hands of people searching for information should likewise increase (rather than diminish) control. Every element of design and technique in the MIR project starts with user needs. In simple terms, people matter. If that sounds like a plea for market-oriented technology, yes, it is!

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

3.2 Coping with data glut

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

People need to know. But facts, or data, are not in themselves knowledge. Facts are like jigsaw puzzle pieces. We must have the pieces, or the puzzle will not come together. And we don't want to miss any relevant facts. But there are too many facts... jigsaw puzzle pieces... that don't contribute to our specific aim at any one point in time. Piling on more and more facts does not necessarily lead to knowledge. Data without recognizable patterns is noise. Noise leads to stress and loss of function. If there is a feeling of being swamped with data, finding desired patterns is all that much harder.

Change has become the norm, change driven by forces such as the proliferation of new products, government regulation, social and technical complexity, communication improvements, customer autonomy, fragmenting markets, and so forth. One notable result: our world is awash in a sea of data. Organizations have to keep track of far more details than ever before. Consider your employer as an example, or any government department with which you are familiar; how much more data is kept today than ten years ago? With few exceptions, you find that there is an exponential explosion of data kept, data required, and information to be retrieved.

Numbers of databases are growing. So is the size of the typical accumulation of data. This is illustrated by what happened in the CD-ROM industry. When compact optical discs were first used for storing computer data in 1985, people wondered how a disc with a capacity of more than half a gigabyte could ever be fully used. Now it is common for a single database to span several CD-ROMs. The cost of new storage technology for personal computers is dropping fast. More storage means more data, and that in turns means an increasing need for quality search capability.

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

3.3 Empowering users

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

"I want what I want when I want it." True for executives. True for two year olds. And, if we care to admit it, true for ourselves when we are searching for information.

Anything can be found, if one has forever to find it. But the average person hasn't got forever. And the time that is available is too precious to be used staring at a "searching database..." message on a non-responsive computer screen. Now even the most amateur retrieval system finds things fast within a small sample (which explains why so many sales demonstrations are done with small samples). More sophisticated textbook indexing methods have acceptable levels of delay for 20,000 (and sometimes even as high as 100,000) records. But today's databases very often exceed these limits. So there has been a shakeout among computer methods of indexing and retrieving information. Only the more powerful techniques of indexing and retrieval can compete on gigabyte-sized tasks.

The primary need is to place a high value on people's time. (So many managers miss this simple truth.)

A second basic need is simplicity. This derives from the need to value the user's time. People do not want to invest time in reading manuals and learning complex systems. Ideally, the searcher should be able to re-use a familiar and preferred search and retrieval system on any new set of data that comes to hand. Maximum gain; minimum pain.

The third need is access. People are empowered to find information as timely data is made available to them at reasonable cost.

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

3.4 Empowering an industry

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

Over the past 30 years, an entire industry has grown up around the requirement to equip persons and/or organizations to extract useful information simply and quickly from quantities of data. The industry launched itself primarily from government data which was distributed on paper, microfilm, microfiche, punched tape, and eventually magnetic tape. By the end of the 1960s, the industry was experimenting with on-line electronic information services. It was the advent of personal computers and optical discs in the early 1980s that made possible information services at dramatically lowered costs. Electronic search split quickly into on-line for the most current data and CD-ROM for historical data. The lowest cost medium has been CD-ROM, which offers the potential for random access across more than 600 million bytes in under two seconds.

The ongoing needs of this industry have to do with development costs, the vast array of formats in which data is received, and disarray with respect to standards.

Development time and costs are much too high because all the better indexing systems have been proprietary. MIR Tutorials and software aim to make world class search and retrieval systems available to the public under the Free Software Foundation "copyleft" rules. Firms may adapt MIR source code for their commercial purposes without payment of license fees or royalties. Costs are also reduced as firms take advantage of automated indexing techniques.

Variability in format of data to be indexed may diminish in the long term, but for at least the remainder of the twentieth century it will continue to be a problem. We address the problem here by offering techniques to cover a wide range of cases.

Standards are an issue because the end user is often forced to learn a new retrieval system when access to a new database is acquired. The current standard for CD-ROM data on compact discs is ISO-9660. This governs how file locations on a compact disc are listed in a directory, but has no bearing whatsoever on the actual content of an index file. Other standards have been developed, for example, Office Document Architecture (ODA), Standard Generalized Mark-Up Language (SGML), etc. These are helpful, but neither do they make it possible to search in uniform ways across totally different databases. CD-RDx and SFQL (Structured Full-Text Query Language) each propose to deal with the problem by engine-independent techniques in which the software dealing with proprietary indexes is separated from the software experienced by the searcher. Each approach has merit; a variation of one or the other may become a new standard in CD-ROM usage. The MIR project aims to facilitate the advance by suggesting index structures compatible with either system. By reducing costs so that many players may use similar structures, and by encouraging improvements and discussion through interactive publishing, we lay the groundwork for development of more extensive standards in the future.

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

3.5 Beyond fast search

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

Automated indexing and full text search of a wide variety of data are, in themselves, immensely worthwhile. The value of this technology goes much further. It serves as a foundation for other possibilities. Among them...

> self-indexing hard disk systems for personal computers;

> correlation software;

> automated detection of trends within a company's production or financial control system;

> records management applications;

> operating systems with indexing power;

> software that learns.

More on these in TUTORIAL FIVE!