

## application note - bioinformatics

*The Human Genome Project catalyzed an explosion in the number of databases to be mined and data types to be analyzed for meaningful molecular biology patterns and rules. Bioinformatics researchers have been overwhelmed by a flood of data and information as high-throughput molecular sequencing and gene expression are accelerating the already exponential growth of bio-information.*



### bio-information explosion

The challenge of keeping up with the latest research confronts the FDA, NCTR, NIH and numerous other public and private agencies that are regulating *and* participating in the drug discovery process, as well as the growing number of companies using bio-information to create new technologies. In the past, print journals were the primary repository of the results of new research. But today, the omnipresent digital domain has enabled new forms of communication that demand intelligent language processing technologies capable of isolating the latest information that is most important to the researcher or regulator.

An objective of bioinformatics is to extract useful knowledge from the flood of data, including biological texts, for the purpose of further analysis leading ultimately to drug discovery; in short, turning the flood of new bio-information into useable knowledge. But the data mining and knowledge management technologies that are being deployed today to assist researchers and regulators are unsuited to this task, for three reasons:

- ▶ **Ever-expanding information** — The results of groundbreaking research are being published every day, at a rate faster than any researcher or regulator can keep up with.
- ▶ **Ever-expanding sources** — To make matters worse, updated bio-information resides in many different locations: print journals, web-based journals, chat rooms, and various intra- and internet data resources. With the sheer amount of new information out there, more than ever it is imperative to return search results that are relevant, so that time is not wasted sorting through irrelevant information.
- ▶ **Bioinformation language** — The language used in bioinformation texts presents unique challenges to any information technology because of its proliferation of special terminology and symbols.
- ▶ **Conventional search technology** — The technology underlying typical knowledge management applications is not up to the task of dealing with these unique challenges, much less recognizing information that is relevant to a particular researcher or research program.

If there were a way to customize a knowledge management application to handle the challenges unique to bio-information texts and return only information relevant to the researcher's interests, then researchers could spend more time applying the knowledge they've gained than on searching for it.

Clearly what is needed is a next-generation technology that is intelligent enough to read unstructured text and return only relevant information. In short, a technology that is intelligent enough to turn simple information into knowledge that researchers can use confidently in the development of new bio-technologies.

## thematrix – the bio-information solution

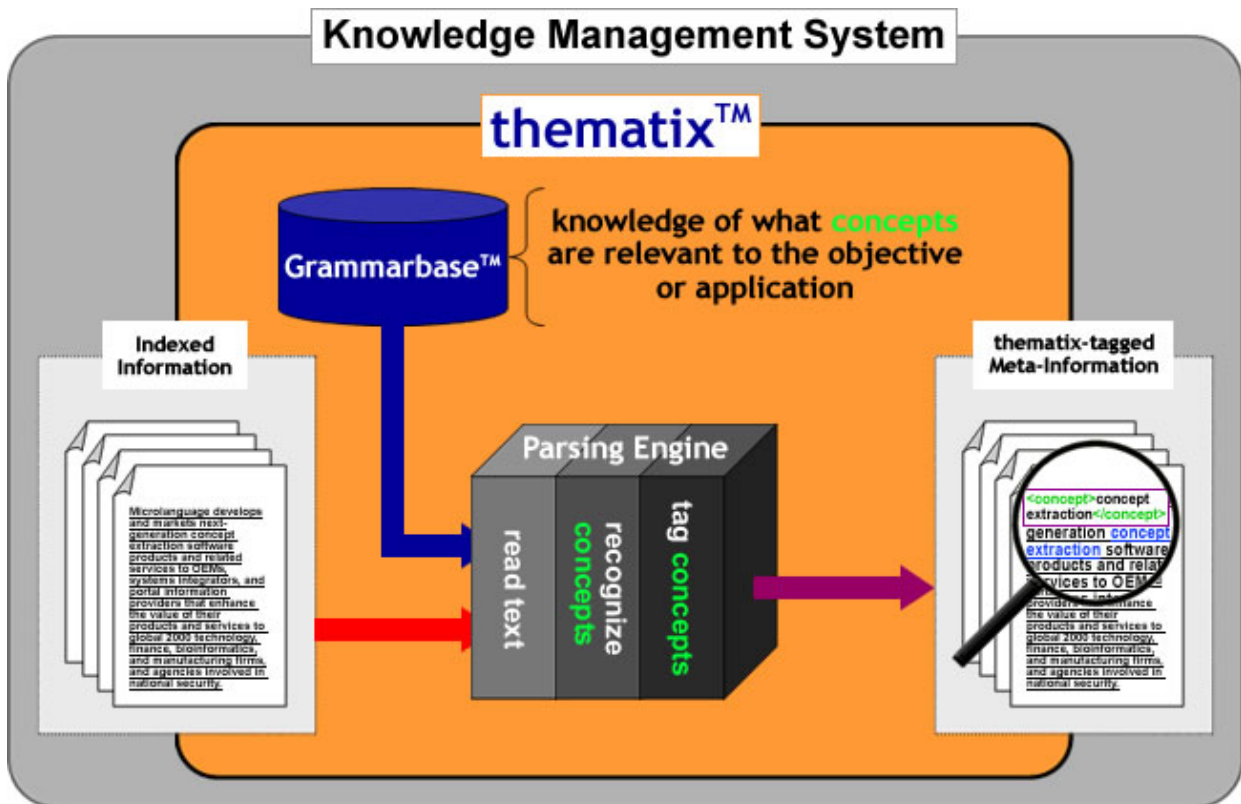
Thematrix is Microlanguage's next-generation language processing technology. It integrates with knowledge management applications to provide intelligent processing of unstructured digital information. Two component technologies work together to create the Thematrix technology:

**Parsing engine** — This component breaks incoming information into its constituent parts, such as words and punctuation so that all the pieces can be examined individually.

**Grammarbase** — This component is a repository of knowledge that gives further instructions to the parser to construct meaning out of the parts.

The knowledge of what information is relevant to the researcher or regulator is programmed into a grammarbase so that Thematrix can find it and they can use it more quickly without wasting time reading through other information that is not relevant to the research objective.

As a component software technology, Thematrix integrates with third party knowledge management applications. Thematrix gives knowledge management applications much higher search accuracy, while interfacing seamlessly with their document management tools.



## benefits of thematix

Thematrix technology enables researchers to focus on subject matter and not on sorting through thousands of reports and communications that may or may not contain information that is relevant to their specific interest. The Thematrix solution offers a variety of benefits:

- ▶ Reduced cost of information processing
- ▶ Reduced researcher fatigue
- ▶ Conserved processing time
- ▶ Closed gap between bio-information and usable information

## about microlanguage

Microlanguage develops and markets next-generation concept extraction software products and related services to OEMs, systems integrators, and portal information providers that enhance the value of their products and services to global 2000 technology, finance, bioinformatics, and manufacturing firms, and agencies involved in national security. For more information on Microlanguage and Thematrix, visit us on the web at [www.microlanguage.com](http://www.microlanguage.com), or call us at 610.995.1017.

Copyright © 2002-3, Microlanguage, Inc., All rights reserved Published March 2003.  
Microlanguage™, Thematrix™, and Grammarbase™ are trademarks of Microlanguage Incorporated. All other product or corporate references may be trademarks or registered trademarks of their respective companies.

