

# Extracting Linguistic DNA

## NStein Goes to Work for UPI

*Bill Trippe*



It's a tantalizing problem for categorization. United Press International (UPI) has more than 700 correspondents creating thousands of stories every week, running the gamut from business news to sports to entertainment to global coverage of America's war on terrorism.

The stories must be filed quickly, categorized, and then made available to thousands of subscriber publications and Web sites. UPI's problem becomes even more tantalizing when you consider its service provides these stories in several languages, and that a key feature of the service is an electronic archive spanning more than 20 years.

For subscribers, effectively combing wire service stories and selecting the appropriate ones is a key to success. Which primary stories should run, and with which sidebars? What stories should be followed and updated throughout the day? What stories and photos would complement some ongoing coverage of a key topic? With hundreds, sometimes even thousands of items to choose from, the manner in which these stories are categorized becomes crucial to the subscribers. How do they understand what is coming in over the wire and how can they best use it?

And while UPI and others news services have mechanisms for adding keywords and categorizing their content, UPI recognized a need to add more automation to the process. With the recent growth and improvement in tools for Computer-Aided Indexing (CAI), UPI undertook a process of looking at its needs and evaluating the many CAI tools out there. In the end, they chose technology from Montreal-based NStein Technologies. "Our main objective was to acquire the best CAI tool to help improve our customers' access and

interaction with our content,” says Steve Sweet, CIO at UPI. “We examined a number of solutions, and NStein’s NServer suite clearly came out on top. The combination of speed, scalability, accuracy, and flexibility was what really sold us.”

### EXTRACTING CONCEPTS

The challenge for an organization like UPI has moved beyond what full-text searching, keyword extraction, and even human categorization can do. CAI tools are able to quickly and accurately make associations and assign categories to content. In NStein’s case, it does this through a focus on concept extraction, an approach where it feels they have a competitive edge. NStein combines statistical processing that provides high-speed access to key terms with linguistic processing that identifies key syntactical and semantic structures in the document. As a result, NStein claims to encode a “Linguistic DNA” (its term) on documents, allowing NStein to provide both the larger categories used in a document along with access to the categories in context. Easier said than done of course, but NStein was able to prove this to UPI.

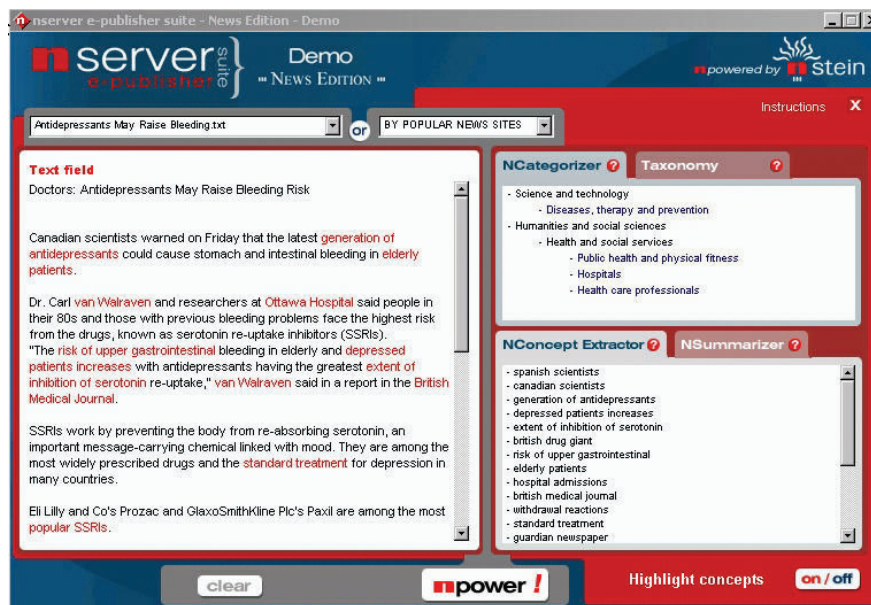
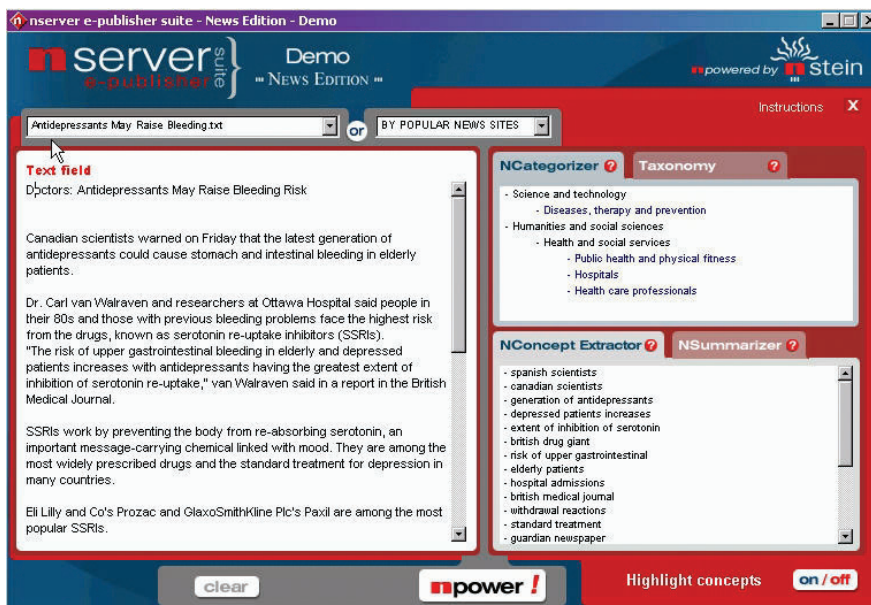
UPI engaged a consulting firm, Marcinko Enterprises (MEI) of San Francisco, to evaluate potential vendors. According to Randall Marcinko, NStein was chosen over its competitors after a detailed and extensive study of twelve companies. Scrutiny was based on parameters that included precision, recall, purchase price, cost of implementation, ease of integration into the UPI workflow, ease of use of the UPI taxonomy, and more. The final group of three vendors was visited and a sample set of 100,000 records was given to each in order to evaluate their product. “After an in-depth study of the results by MEI lexicographers,” says Marcinko, “NStein was found to be the clear winner.”

As Charles Alexander, NStein’s vice president for marketing and business development explains, “Linguistic DNA technology extracts concepts from texts to determine a document’s linguistic fingerprint.” The engine compares these concepts

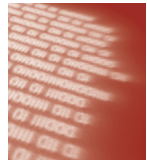
against a knowledge base and then matches them to defined categories. Because it focuses on multiword units rather than keywords, says Alexander, “Our linguistic DNA technology produces much more meaningful results than keyword-driven technologies. Over time, the knowledge base grows constantly and learns by adding the new concepts to the knowledge base—

automatically creating new associations and providing smarter results.”

So how is UPI actually implementing and using the NStein technology? According to UPI’s CIO Sweet, UPI will utilize the NStein tools to produce part of the metadata for any item that includes text, be it article text, caption text, or the text in audio scripts, video scripts, and transcripts. To



The information specialist is presented with candidate concepts that have been extracted from a training set of documents. Using a visual interface, the information specialist can readily accept, deny, or refine the candidate concepts. This interaction between the automated tool and the information professional is at the heart of computer-aided indexing.



The NStein technology fits squarely in UPI's workflow, which for years has included processes for coding content according to a system from the International Press and Telecommunications Council.

speed up the taxonomy work, UPI has asked Marcinko Enterprises, which helped evaluate NStein, to create UPI's News Taxonomy. Information specialists from Marcinko used the NStein software to extract concepts and assist in the development of the taxonomy.

According to Sweet, UPI's use of NStein is part of a bigger, ambitious project to both revamp its taxonomy and upgrade its multilingual editorial system. "UPI is anticipating availability of the first English public beta during the first quarter of 2002," says Sweet. The rollout of the NStein-supported taxonomy is part of a two-year infrastructure project at UPI. "The English language component has been deployed, and UPI is currently developing the Arabic and Spanish language components for integration," reports Sweet, and additional language components will be developed as UPI expands. Impressively, as UPI expands, Sweet explains, "language-specific news taxonomies will be developed and integrated with the English language taxonomy for cross-reference and relevancy."

NStein's product is Nmedia Server, which consists of four independent modules:

*The **concept extractor** locates and isolates concepts and meanings embedded in documents. It also identifies concepts not defined in categories without the aid of a*

*preset thesaurus and uncovers significant connections.*

*The **categorizer** uses text-parsing techniques and sorts content against standard or customized taxonomies. The module categorizes and classifies based on concepts rather than the number of times words are used in given documents*

*The **tagger** recognizes and labels words or groups of words and automatically generates XML tags for them. This helps ensure that the processed files are interoperable with other XML-compliant database systems.*

*The **summarizer** provides summaries of documents and subdocuments, and can assign categories to the summaries.*

The NStein technology fits squarely in UPI's workflow, which for years has included processes for coding content according to a system from the International Press and Telecommunications Council (IPTC). New content flows through the desk editors, who assign codes to indicate the subject matter, relevancy, priority, and type of article. The NStein tagging is then applied to the finished, coded content, prior to it being packaged and delivered to UPI's subscribers.


Existing content, including UPI's archive of thousands of stories, is stored as fielded data in a SQLserver database. NStein is then used to index the database, to good results. For subscribers who then receive access to either new content or the archives, their feed will now include enriched IPTC/NewsML metatagging to go along with the text itself.

## **A SIGNIFICANT PARTNER FOR NSTEIN**

As important as this project is for UPI, it is even more significant for NStein, a relatively new player in the CAI space. NStein Technologies is a public company listed on the Montreal Exchange since June 2000. The technology was originally developed by computational linguists at Laval University, and NStein was formed to bring the technology to the marketplace. The UPI business

helps NStein in a number of ways—it's a large deal in and of itself, but it also gives NStein a blue-chip reference customer and a beachhead in the United States market. The agreement to license NStein's technology includes the base license in English, French and Spanish, as well as options for modules for Korean, Japanese, Arabic, Russian, German, and Chinese language modules. NStein is also providing integration, training, and consulting services.

From UPI's perspective, the collaboration with NStein has been a complete success so far. Asked if there were any negative surprises in the implementation, UPI's CIO Sweet said, "So far surprise has kept itself away. The ongoing development and relationship with NStein has been superlative." Consultant Marcinko had a similar assessment, saying the implementation has uncovered no negative surprises. "Positively, we have been very pleased with NStein's capabilities to do entity extraction and to generate both extracted concepts and associated categories from the UPI taxonomy," said Marcinko. "The NStein linguistic DNA has provided a tool for our staff to use in the creation of a 'more-like-this' feature for the UPI site."

Significantly, it is this kind of feature that UPI, and its subscribers, have been looking for, and that CAI technology, properly used, can help bring to market. For NStein's Alexander, this is exactly the kind of thing he sees his technology being best applied to. "It's where the consistency of computers and the quality of human beings can be best combined," says Alexander. And what's exciting for Alexander, and for others involved in the CAI world, is that these kinds of applications are now not only possible, but in fact are happening. "What didn't work before, works now," says Alexander, "and the possible applications are endless." 

**BILL TRIPPE** (btrippe@nmpub.com) is the founder of Boston, Massachusetts-based consulting practice New Millennium Publishing. Comments? Email letters to the editor to [ecletters@onlineinc.com](mailto:ecletters@onlineinc.com)

**n.**power

**|||** the power to do what you do. **better.**

**n** server suite  
e-publisher }

Introducing the most advanced content management indexing software. The nserver suite by Nstein. It's faster, more precise and more flexible than any other indexing software on the market today. Let us show you how we can help you maximize the value of your content and npower the way you do business. Call today 1.877.678.3461 or visit [www.nstein.com](http://www.nstein.com).

**npower your categorization • npower your retrieval • npower your decisions**

