# IBM to Open Source Technology for Analysis of Unstructured Information

August 8 2005

IBM announced plans to make available through open source its Unstructured Information Management Architecture (UIMA), a technology designed to support a new breed of software applications that can process text within documents and other content sources to understand the latent meaning, relationship and relevant facts buried within.

There has been an explosion of "unstructured" information in the enterprise and across the web, taking the form of documents, images, comment and note fields, e-mail and even rich media like video and audio. However, the technology has not existed to allow software to search out and make sense of these disparate forms of data. UIMA provides an open software framework with standard interfaces for adding unstructured information analytics to any application. This framework makes it easy to integrate the analytic software tools and end-to-end enterprise applications across several different vendors. UIMA also provides tools to speed the creation of new, reusable analytic software components to handle unstructured information.

The result of more than 4 years of development by IBM Research, UIMA has also received significant support from the Defense Advanced Research Projects Agency (DARPA), the central research and development organization for the Department of Defense. DARPA and IBM formed a working group consisting of experienced research members who have contributed their expertise in unstructured information management to the evolution of UIMA.

"DARPA is always interested in computing infrastructure that might allow significant numbers of people to leverage their work -- you can look at the ARPANet and the Internet, both of which came out of our office, as examples. Projects like these allow the government to achieve a large, multiplicative effect with only a small investment of money," said Dr. Ronald J. Brachman, Director of DARPA's Information Processing Technology Office. "We saw the potential of UIMA to help bring together and multiply the effects of the work of a large community of researchers, and are excited to see the strong momentum and support shown by the Working Group members. Having an open-source framework for deploying text analysis components will help us deliver more advanced solutions for the national security community."

The contributors included several leading universities, along with industrial research and development organizations. Some of the universities that participated, such as Carnegie Mellon University, Columbia University, Stanford University and The University of Massachusetts Amherst, are already using UIMA in courses and research projects. The other organizations actively supporting and using UIMA include Science Applications International Corp., BBN Technologies, The Mayo Clinic and MITRE Corporation. In addition, widespread commercial adoption of UIMA was announced today among more than 15 software vendors.

"UIMA provides, for the first time, true interoperability among different knowledge discovery, search, business intelligence and text analytics software," noted Arthur Ciccolo, Department Group Manager for Information and Knowledge Management, IBM Research. "This initiative will enable organizations to deliver groundbreaking solutions that can leverage unstructured information in entirely new and advanced ways."

This technology can be leveraged by advanced knowledge discovery and

business intelligence applications to incorporate information that was previously difficult to understand and leverage. UIMA can also be used to support broader enterprise search applications by providing a common mechanism for developing and delivering natural language processing solutions. It can help computers understand and reason across the wide variety of information sources available to humans. Ultimately it will enable computers to extract deeper levels of meaning from such information -- including the relationships that define specific facts.

The UIMA framework has already been embedded in IBM products, including IBM WebSphere Information Integrator OmniFind Edition, the first commercially available software platform for processing content based on the UIMA standard. IBM WebSphere Portal Server and Lotus Work Place also leverage UIMA for content processing.

The technology will be presented to Open Source Technology Group with availability through SourceForge expected by the end of year 2005. The UIMA framework can currently be downloaded free of charge from IBM AlphaWorks at www.alphaworks.ibm.com/tech/uima .