

# A MULTIMEDIA SYSTEMS APPROACH TO NATIONAL SECURITY POLICY, DECISION MAKING, AND INTELLIGENCE SUPPORT

*John P. Crecine and Michael D. Salomone*

The amount of raw data available to military and national intelligence professionals is growing exponentially. The revolution in our ability to generate and disseminate data has not led to a corresponding increase in our ability to convert data to useful information. The revolution in information technology, coupled with the natural tendencies in military and intelligence organizations to compartmentalize data (especially when collected using automated or covert methods), combine to hinder military commanders, top-level analysts, and national decision makers from coherently interpreting this profusion of information. The “stove-piping” of data within subunits of an organization makes effective integration at the top much more difficult.

Existing data processing systems make some kinds of information extremely useful, but the systems are incapable of storing, organizing, processing, and retrieving other data types, including text and images, in ways that allow top-level decision makers and analysts to make coherent, comprehensive use of the information. Systems based on new technologies can enable these professionals to more efficiently manage data and information available in electronic forms. The combination of an object-relational database management system with a natural-language processing-based, free-form text search engine, paired with image recognition software and standard structured query language-based search, combined with recent advances in processing, storage, display, and telecommunications technologies, holds great promise for effective information management systems of the future.

Intelligence is information. Raw data from multiple sources and in a variety of forms, brought together in a

coherent and timely way, integrated and congealed in the context of particular policy and operational issues or in the

interest of a comprehensive interpretation of particular events, is transformed into useful information. The challenge of managing data available in digital and electronic form, as it is transformed into useful information, is enormous—both in terms of the vast quantities of data and in terms of the incredible diversity and complexity of the raw data elements routinely collected. The fundamental problem is that a huge, eclectic, and rich array of data must be organized and stored in such a way as to allow for retrieval in targeted and often unanticipated ways for unanticipated purposes.

The net effect of modern information and communications technologies has been to exacerbate rather than alleviate the information management problem for all professionals, not just those in the military and intelligence communities. Compartmentalization and division of labor,

the hallmark of organization and bureaucracy, creates dedicated, specialized, and separate information management and data processing systems. This “stove-piping” effect means that different data types and data collection organizations have a tendency to operate as closed systems, up through several organizational levels. Once data has been processed through several organizational levels, the initial data elements or observations or transactions are “lost” or unavailable to top-level analysts or decision makers. Whether the initial data element is a paragraph in a trade journal or an internal government memo, or is a portion of the yield from a satellite’s remote sensor, a readiness report for a low-level military unit, or a simple accounting transaction, the raw data is seldom easily accessible to top-level analysts or decision makers, and it is almost always impossible to combine

**Dr. John P. Crecine** earned a bachelor’s degree in industrial management, and a master’s and doctoral degrees in industrial administration at Carnegie Mellon University. At the University of Michigan, he established the country’s first graduate program in public policy in 1968. He subsequently served as dean of the College of Humanities and Social Sciences at Carnegie Mellon and later was appointed Senior Vice President and Provost. He oversaw Carnegie Mellon’s academic, research, and systems development in computing and computer science and initiated the formation of the School of Computational Sciences. In 1987 Dr. Crecine became Georgia Tech’s ninth president. During his tenure, he established three new Colleges at Tech—the College of Computing (the first such college in the country); the Ivan Allen College of Management, Policy, and International Affairs; and the College of Sciences. After Dr. Crecine’s resignation in 1994 as President at Georgia Tech, he became involved in the development of educational and business computer software.

**Dr. Michael D. Salomone** is Professor of International Affairs in the Sam Nunn School of International Affairs, Georgia Institute of Technology, Atlanta, GA. He was educated at Lehigh University and the University of Pittsburgh, and has served on the faculties of Bethany College, the University of Pittsburgh, and Carnegie Mellon University prior to joining the faculty of Georgia Tech. Dr. Salomone is a consultant to numerous departments and agencies of the United States Government on issues of defense policy, strategic planning, and international security, and was a member of the professional staff of the U.S. General Accounting Office, International Division. He is a member of the International Institute of Strategic Studies in London, a Fellow of the InterUniversity Seminar on Armed Forces and Society, a Fellow of Sigma Xi The Scientific Research Society, and a NATO Institutional Research Fellow.

data from different sources (“stovepipes”) to suit unanticipated needs. Modern information technology generally accentuates and multiplies the stove-piping problem.

As we move into the next century with an accelerating pace of technological change, an important question is whether modern information and communications technologies can be used to alleviate a problem they have helped create. Can technology aid in the targeted retrieval and purposeful synthesis of the technology-driven explosion in the quantity and diversity of raw data in analog and digital formats? Just as it took an important enabling technology—HTML (HyperText Markup Language), Web browsers (Mosaic/Netscape), and search engines (Lycos, Excite, AltaVista, Yahoo)—to make the Internet useful for tens of millions of users with eclectic interests, military, intelligence and other professionals need enabling technologies to make the rapidly growing mountains of increasingly diverse data types professionally purposeful.

In the language of the national security and intelligence community, “Can modern information technology rescue the intelligence fusion process from the enormous information overload, storage, retrieval, and synthesis problem?”

## **THE NATURE OF THE DATA AND INFORMATION MANAGEMENT PROBLEM**

---

This challenge is complicated, since data relevant to a particular issue is collected from a very wide array of sources, usually by different people in different organizations, and often for narrower purposes than those of relevance to national intelligence organizations, policy makers,

or a military commander. The efficient and cost-effective management of potentially relevant data and information is a central issue for all organizations, and particularly so for the top-level decision making, policy, and intelligence functions in military organizations and the national intelligence community. Data and information relevant to these functions is incredibly diverse, compiled in great quantities, in widely varying formats, and from wide-ranging and shifting numbers of sources.

“ The efficient and cost-effective management of potentially relevant data and information is a central issue for all organizations...”

Much of the data waiting to be collated and synthesized into information are memos, journal articles, news accounts from the print media, reprints, translations, notes, papers, and transcripts—free-form text. Still other data take the form of photo images, sensor outputs, signals, video images, tabular alpha-numeric data, maps, news accounts from the electronic media, or audio clips. Data and information resources are generally dispersed geographically and organizationally. The management of information for top-level decision making, policy, and intelligence functions is a formidable task and, as currently conducted, is necessarily very labor intensive.

Consider only the accelerating information management sub-task of battlefield management. As new sensors and intelligence gathering devices are created to support development of dominant battle space awareness, there is a corresponding and geometric increase in the amount of information the commander and the supporting intelligence organizations must

comprehend, at least in theory. Effective commanders must be increasingly selective in the information to which they attend, or they need to capture the benefits of automation in executing the information management portion of their command function.

Consider the real-time intelligence analysts' task of interpreting rapidly changing events, with a pile of articles, notes, demographic charts, biographical sheets, maps, books, Cable News Network (CNN) video tapes, and newspaper clippings on his or her desk.

"In a bureaucratic setting—be it a military or intelligence organization—information seldom is delivered to a commander or analyst without an associated context, framework, or interpretive suggestion."

First, is this eclectic pile of data the right pile for this event? Second, can the analyst quickly sort through, analyze, cross-reference, and correlate information in support of a competent situational analysis? Under the best of circumstances, this a daunting information storage, retrieval, and organizational task.

The problems faced by the commander in a battlefield situation or by an intelligence analyst are surprisingly similar. Due to the pressures of time or the overwhelming amounts of potentially relevant information, there is an information overload problem of considerable proportions. Whether individuals are seeking the "right answer," an accurate situational analysis, an adequate understanding of events, or a

framework for understanding the phenomena or event of interest, individuals must be selective, focused, and guided by some sort of paradigm or crude framework in interpreting the data they have. For the most part, history and experience, shaped by training, dominant scenarios, and "lessons learned" from roughly similar situations, help individuals sort through the infinite combinations and permutations of potentially relevant information.<sup>1</sup>

In a bureaucratic setting—be it a military or intelligence organization—information seldom is delivered to a commander or analyst without an associated context, framework, or interpretive suggestion. Often organizational subunits are formed around a particular paradigm or belief system and information coming from such organizational units comes laden with those biases and selectivity filters. Organizational preprocessing is hardly unexpected. The various elements of the technology-driven information revolution combine to deliver vastly greater quantities of information in vastly more diverse formats (data types), resulting in a geometric growth in the potential information processing and synthesis problem. This problem is made manageable in current settings by even greater reliance on dominant paradigms generated by personal histories and training and by even greater selectivity and filtering by bureaucratic information providers. The information revolution has been, in fact, a revolution in data generation and dissemination, not in data understanding.

The information revolution does not naturally create a greater ability of chief

<sup>1</sup> This is illustrated in Graham T. Allison's classic *Essence of Decision*, (Little Brown, 1969) in analyzing the Cuban Missile Crisis using three theoretical approaches or points of view.

executive officers, commanders, and analysts to “get it right.” Here we’ll discuss some of the technical reasons why the revolution in our ability to generate and disseminate data has not led to a corresponding increase in our ability to convert data to useful information. We’ll describe some technical developments that we believe can help improve the data-understanding-to-useful-information conversion process, through a partial automation of the new, more diverse, and difficult information management task.

The explosion in the volume and diversity of intelligence information is occurring in an environment of declining financial resources for intelligence assessment, military decision making, and national security policy making. Professionals in these areas must achieve dramatic increases in the efficiency and effectiveness of their information management and fusion systems, merely to maintain current levels of effectiveness of intelligence, decision making, and policy functions.

### **APPLYING RECENT ADVANCES IN TECHNOLOGY TO THE CHALLENGE**

---

How might one bring modern information processing technology to bear on national intelligence tasks? Recent advances in many areas foster the hope of creating affordable, comprehensive information systems. Among these are:

- multimedia-data applications,
- analog-to-digital conversions and data compression,

- artificial intelligence applications in image recognition and natural language processing,
- modern telecommunications, including both wide-area and local-area network technologies,
- client/server computing systems,
- distributed database systems,
- visualization and display technologies,
- digital storage media, and
- computing technologies.

Systems based on new technologies can enable government and military personnel to more efficiently manage the exponential growth rate of data and information available in electronic forms, and to more effectively deal with the challenges presented by the proliferation of data formats. A comprehensive and affordable approach to the intelligence information management task is now possible due to advances on several technical fronts.

“Major advances have occurred recently in the ability to store, access, and process vast amounts of raw data in a variety of forms, and to do so using digital data formats.”

Major advances have occurred recently in the ability to store, access, and process vast amounts of raw data in a variety of forms, and to do so using digital data formats. First, recent advances in optical and magnetic technologies make the storage and retrieval of terabytes (a

million megabytes) of digital information possible and affordable. A “jukebox” of high-capacity, read-write compact disks now costs less than \$5,000 and stores a terabyte of information. These new jukeboxes can replace a \$2 million to \$3 mil-

“...it is the advance in information storage and retrieval strategies (database management systems, or DBMSs) that makes it possible for professionals to facilitate the appropriate collection of needles in an expanding set of exponentially growing, data haystacks.”

lion system that was state of the art only one or two years ago. And the new, digital video disks (with an order of magnitude greater capacity) are not far behind. In order to grasp the magnitude of this available storage capac-

ity, a terabyte represents more text than is stored in most university libraries.

Second, rapid advances in video compression technologies, driven by the broadcast entertainment industry, have resulted in the electronic storage and retrieval of complex, high-resolution digital images, full-motion high definition television (HDTV), and animations and simulations. Data compression and expansion involves conversion to a digital format, which increases the accuracy and reliability of communication transmissions and provides for a greater ability to manipulate sensor and image data.

Third, the super computers of a couple of years ago are now the reduced instruction set computing (RISC) chips in personal, desktop computers, at 2–5 percent of the cost. Processing power for digital data formats is available, increasingly powerful, and inexpensive. Parallel

processing PCs are now readily available for \$5,000 to \$10,000 (usually configured for image processing and multimedia production tasks).

Finally, the information superhighway, fiber optic cabling systems, digital direct broadcasts, cable, and satellite systems coupled with advances in client/server architectures all mean that the collection and storage of information can be geographically and organizationally distributed, without paying the usual penalties (e.g., lack of access or incompatibilities).

The rate of change in technology is so rapid, the efficiencies to be gained so great, and the cost savings so massive, that to use anything other than commercial, off-the-shelf (COTS) hardware and software would be foolhardy. An acquisition process that fails to closely track the massive, rapid, and revolutionary advances of the commercial sectors of modern information technology is fatally flawed.

For these advances, largely in hardware and microelectronics, to lead to real advances in the information management side of military and intelligence functions, there must be corresponding advances on the information processing or software portions of the information management system.

---

### **ENABLING TECHNOLOGIES: DATABASE MANAGEMENT AND THE PROBLEMS OF FREE-FORM TEXT AND IMAGES**

---

Although advances in hardware, storage, networking, and display technologies make many storage and retrieval tasks economically and technically feasible, it is the advance in information storage and

retrieval strategies (database management systems, or DBMSs) that makes it possible for professionals to facilitate the appropriate collection of needles in an expanding set of exponentially growing, data haystacks. Data are generally in locations remote from users. Different databases assembled, owned, and maintained by different agencies and sources can be distributed over a secure network. Distributed databases can be comprehensively searched and information in multiple forms can be selectively retrieved and brought to bear on questions of particular interest to analysts or policy makers in a timely fashion.

### **REQUIREMENTS FOR ACCESS TO USEFUL INFORMATION: DATABASE SYSTEM AND SEARCH ENGINE**

---

Two new areas of progress combine to make possible a broad, general purpose national security and intelligence information management system. One area consists of a new approach to database management: a data storage and organization strategy. For example, existing relational database management systems (RDBMSs) allow the user to identify books in a database, based on a Library of Congress code or index. A new, object-oriented database management system (ODBMS) allows a user to examine not only the Library of Congress code, but all of the text in all the books in the database and go straight to the relevant passage or paragraph.

The second area relates to “search engines” for free-form text and images, a data search and retrieval strategy. This consists of, first, a natural language processing

“search engine,” where concepts and abstract representations of content, as well as key words, are used to search and identify relevant information (the needles) among the wide array of data types (the haystacks), and in a far more precise manner than has ever been possible. Second, it includes image recognition software that allows one to search digital

“For more than 25 years, relational databases have formed the basis for most of the automation of administrative and financial information systems.”

images (still or video) based on color, composition, objects, structure, or other characteristics to identify particular records that correspond to search criteria or user-provided examples.

### **RELATIONAL DATABASE MANAGEMENT SYSTEMS AND THEIR LIMITATIONS**

---

For more than 25 years, relational databases have formed the basis for most of the automation of administrative and financial information systems. Relational databases allow multiple users to systematically search a range of different, fixed-record-length files. Fixed record length means relational databases work very well for numbers and alphanumeric data of fixed length (e.g., a 35-character record containing formal names). In the search and retrieval process, the RDBMS can examine individual records using a standard query format (SQL, or structured query language, the industry standard), and pick out those records that meet the criteria.

When data is located in complex multimedia formats (e.g., a journal or news article or free-form text, a video, an image, a computer-aided design [CAD] drawing, or an audio recording), traditional relational databases store data elements as uninterpretable BLOBs (Binary Large Objects). Customarily, relational database users will develop a coding scheme or “tag” for each type of BLOB. The data or content in any particular type of BLOB is accessible only through the coding scheme. For example, the text and content of a library book is stored as a BLOB in a RDBMS, and the BLOB is accessed through the Library of Congress code or index. The establishment of a fixed coding scheme, in effect, implies that all future uses for the content of a BLOB data

“Moreover, and most important, the law required that an initial budget be produced by an “independent group of recognized experts,” hence the assignment of the job to a team of expert budget consultants (all employed by a prestigious consulting firm).”

type are known and have been embedded in the code. In the book-in-a-library example, the implication is that the Library of Congress code for the book captures all of the relevant content of the book. This assumption is seldom

appropriate for the needs of top-level analysts, decision makers, and policy makers, and is almost certainly inappropriate for any knowledge domain that is rapidly changing. Coding schemes in a RDBMS are difficult to extend, once established.

In response to the development of object-oriented programming and database

technologies, most leading RDBMS suppliers have added what they term as an “object layer” to their RDBMS. The object layer generally refers to a coding scheme used to provide access to data stored as BLOBs.

---

## **OBJECT-ORIENTED DATABASE MANAGEMENT SYSTEMS: OVERCOMING RDBMS LIMITATIONS**

---

In terms of data complexity, relational databases only allow for the efficient storage and retrieval of fixed-record-length, alphanumeric data. The proliferation of rich, and more complex digital data types (non-fixed length records, containing data other than alphanumeric) led to the development in the early 1980s of object-oriented DBMSs (ODBMSs). While allowing for more complex data types, and being entirely extensible with respect to data types, ODBMSs do not have an effective method for querying the data—there is no SQL equivalent.

However, there are many additional benefits of ODBMS that have to do with systems development issues, which use the inheritance and encapsulation nature of “objects” in object-oriented programs. The elements of object-oriented database management systems represent very efficient building blocks for other applications.

---

## **INEFFICIENT METHOD OF COMBINING THE ADVANTAGES OF RDBMS AND ODBMS**

---

Most relational database management systems have attempted to capture the benefits of object-oriented database

management systems by placing an object layer on their existing systems. Although such an approach preserves legacy applications (SQL-based), it possesses significant flaws. Grafting a fundamentally different system to an old architecture means that the basic search engine is unable to understand how to optimize storage and retrieval features of object data records. At best such an approach will be extremely inefficient: slow, and requiring extra processing, storage, and communications resources. The magnitude of the inefficiency problem is geometrically proportional to data complexity and the size of the database(s).

### **OBJECT-RELATIONAL DATABASE MANAGEMENT SYSTEM**

---

A more straightforward and promising approach is to re-architect a DBMS from the ground up, using the following methods:

- Incorporate artificial intelligence software with a knowledge of objects.
- Create “smart objects.”
- Construct query languages (including SQL legacy systems as a subset).
- Redevelop client/server architectures.
- Integrate extensibility of data types into the re-architected system from the beginning.

Such an object-relational database management system (ORDBMS) was created in the early 1990s and has been operationally tested in a wide variety of applications,

with extremely diverse sets of data types. The ORDBMS approach is ideally suited to the size, complexity, extensibility requirements, and the need for flexible (and inherently unpredictable) search strategies that characterize national intelligence database applications.

### **SEARCH ENGINES**

---

Finally, a key to a superior solution to the management of intelligence information is the ability to precisely identify individual records

(data objects) of extremely varied types, and to collate them from a widely distributed set of heterogeneous sources. Natural language has evolved over many millennia as a

way to describe, with precision, diverse kinds of objects, information, and abstractions. Natural language is adapted to the interpretation, perception, and correlation tasks of natural intelligence. A natural language processing search engine is a manageable approach to finding the needles in the diverse, large, and expanding haystacks of potentially relevant information.

Pure keyword indices, whether weighted or not, can provide a useful first cut in identifying and retrieving user-defined, relevant intelligence information. However, the failure to “understand” the

“...a key to a superior solution to the management of intelligence information is the ability to precisely identify individual records (data objects) of extremely varied types, and to collate them from a widely distributed set of heterogeneous sources.”

content of individual data records with no more depth than a simple frequency of keyword mentions in a data record highlights two problems. First, too many “rel-

“In most information storage and retrieval systems...for performance purposes, search and retrieval is a two-stage process.”

evant” records will be identified. Second, if a particular concept or substantive element of content is described in different words in

several different sources or records, many relevant records will be missed. The proposed natural language search engine has the ability to represent a particular unit of text or language as an abstraction of data that contains concepts and information, in addition to including a more sophisticated method for counting keywords within that unit of text.

## GENERAL INDEXING AND SEARCH, AND STORAGE AND RETRIEVAL STRATEGIES

---

### INDEXING AND SEARCH ENGINES

In most information storage and retrieval systems, including several popular Internet browsers (retrieval systems for Web pages), for performance purposes, search and retrieval is a two-stage process. First, all relevant records and data elements in storage are “read and analyzed” by an application or routine that could be viewed as a general purpose search engine. The routine analyzes each record or file

and notices key features of that file, using those features to create an index for that file or record. For free-form text, the index might consist of key words, perhaps weighted by their frequency or relative location in the text. In more sophisticated natural language processing routines, “concepts” (key words, linked together in a particular linguistic structure), and weights attached to the relative importance of key words and concepts in the text might also be included.<sup>2</sup>

The second stage is to retrieve a targeted subset of the records represented by the master index. To do this, the routines analyze a query by the same routines in the same way that individual records were processed in creating the master index, and a special, much smaller index is created for the query. The index for the query is then compared with the master index. The result is a detailed report of pointers to that subset of individual records where there are matches between the query’s index and the master index. Armed with this report, the user retrieves those records that look particularly interesting. The initial indexing of individual data records and files, and the search for and identification of individual records is a very similar process.

In an image understanding routine, visual content—the structure, composition, texture, color, or object—along with ancillary information (e.g., the source or date) might form the basis for the image’s index. A master index could then be constructed that would be an overlapping set or a composite of all the individual indexes for individual records. Nearly all existing

---

<sup>2</sup> The leading knowledge-based, natural language processing system comes from the Center for Machine Translation at Carnegie Mellon University (Jaime Carbonell, founding director). A simpler version of the core of their machine translation system is the basis of the Lycos™ Web browser.

image and video search and retrieval systems rely entirely on narrative, linguistic descriptions to identify a set of images or video sequences that are then examined visually by the user.

### **USE OF LINGUISTIC CONTENT TO IDENTIFY IMAGES**

Indexing and retrieval of video sequences or “clips” from vast stores of video data, based on content, implies that the content of the video clip is used, directly or indirectly, as the basis for storage and retrieval. In order to capture as much content as possible, the linguistic content of the audio track is treated as part of the video clip’s content, and in many cases, for storage and retrieval purposes the audio track content is assumed to capture all of the video clip’s content.

A leading technology in this regard is embodied in the Carnegie Mellon University Informedia Project. Speech recognition software, Sphinx II, is used to automatically translate a video clip’s audio track into free-form text.

Language processing software similar to that used in Lycos™ (<http://www.lycos.com/>), a popular Internet Web page search engine, generates a content-based index for the video clip. It is this index that forms the basis for search and retrieval of video clips. It should be noted that approaches that depend exclusively on natural language understanding generally fail to sufficiently categorize the associated imagery.

Indexes and searches in the approach advocated here involve both natural language processing and image recognition software, functioning both as indexes and search engine.

### **COMPOSITE SEARCH ENGINES**

Search engines designed for use with a single data type (e.g., text, image, video clip, or sensor output) may be useful in delimiting human search of more complex or ambiguous data and content. The question remains as to whether one can make several search engines (one for each data type) work in tandem to produce results superior to one search engine working

“ Search engines designed for use with a single data type... may be useful in delimiting human search of more complex or ambiguous data and content. ”

alone (Jennings, 1994). We anticipate that the answer is “yes.” An important step in creating the type of flexible information management system advocated here for national security and intelligence use is to create a superior, composite search engine from several “single data type” search and retrieval systems. A composite search engine would, at minimum, include both natural language understanding-based and image understanding-based search engines.

Two issues, both heavily influenced by the context and knowledge domain of the search, must be addressed as this composite search tool is developed. First is the role of ancillary information and data in identifying appropriate records. (For example, a military analyst might profitably use historical time lines, place names, person name files, and newspaper articles to correctly locate video clips with particular content.) Second is the purpose of the search, which will help determine relative weights placed on the dimensions of the

search. Both issues influence the choice of the archive of video, text, and other information to test various components of any composite search engine.

### **COMMON APPROACHES TO STORAGE AND RETRIEVAL OF VIDEO SEQUENCES**

The approach taken by the Carnegie Mellon Informedia™ project (mentioned above) to storing and retrieving video material, although considerably more sophisticated,<sup>3</sup> is similar in broad outlines to that employed by large broadcast news organizations. Both make use of linguistic content descriptions for indexing or cataloging video sequences and digital images and for retrieving them.

CNN has developed a system for storing, cataloging, and sometimes retrieving the hundreds of thousands of hours of raw video feeds (eight feeds, more or less continuously, over

“None of the language or image processing components that make up the composite information retrieval and search engine will be 100 percent accurate.”

each 24-hour period) and the 5 percent that make broadcasts.<sup>4</sup> CNN has a human viewer who, as the video feeds are received from the field, di-

vides them into sequences or “clips” with a beginning and end, and types brief notes describing their contents. It is the notes that are searched (not the images themselves), via computer, pointing to a place

on a particular video tape, which has been catalogued, much like a book, and placed on shelves.

That this imagery and its associated linguistic descriptions or narratives/transcripts have only somewhat overlapping content is demonstrated by the fact that more than 80 percent of the requests for historical footage or video clips made to the CNN Library are simply for images or particular kinds of visual scenes and settings. Particular requests for historical video clips generally have little or nothing to do with story content. What is requested is a visual backdrop to an often unrelated story. This approach is insufficient for intelligent analysts and national security policy makers.

### **PERFORMANCE CHARACTERISTICS OF COMPOSITE SEARCH ENGINES**

---

None of the language or image processing components that make up the composite information retrieval and search engine will be 100 percent accurate. An important question (once the “best of breed” search engine components are identified) is to how reliably each component has to perform in order to be good enough. Because each component is embedded in a complex system and interacts with a particular knowledge domain, there is no easy answer. And the best systems will exploit the characteristics of particular knowledge domains.

---

<sup>3</sup> This strategy uses audio track and speech recognition software to generate an automatic transcript that presumes to describe content of the video.

<sup>4</sup> This description results from extensive discussions with Cable News Network, particularly with the News Tape Librarian.

For example, the best available speech recognition software today may be 85 percent accurate in speaker-independent settings with a 20,000 word vocabulary, and 98 percent accurate with an 80,000 word vocabulary, but the larger vocabulary system may cost 16 times as much and take 10 times as long to search. Video sequences indexed using the audio sound track and 85 percent speech recognition may be “good enough”—accurately indexed and retrieved under such conditions with an elaborate, detailed query, which depends on the uniqueness of the clip or query and the redundancy of the responses to it. Or, when coupled with image recognition software that is 75 percent accurate on identifying video clips, the combination of search tools may be such that the accuracy of the composite system is substantially better than the accuracy of either component. Or the opposite may be true, depending upon the content and the characteristics of the knowledge domain. Information on troop movements and real time deployment is different from information about national political subgroups, which is different from water resource information or the characteristics of a nation’s secondary educational systems.

### **ONE APPROACH TO INFORMATION MANAGEMENT FOR INTELLIGENCE ANALYSTS**

---

The combination of an ORDBMS with a natural-language-processing-based,

free-form text search engine paired with image recognition software and standard SQL-based search, when coupled with recent advances in processing, storage, display, and telecommunications technologies, holds great promise for the information management systems appropriate for military and national intelligence functions.<sup>5</sup>

Multimedia Archival Systems, Inc., (MmAS) is developing a robust information management system<sup>6</sup> to accommodate the automatic

indexing, archiving and retrieval needs of intelligence analysts and national security policy makers and decision makers with diverse data needs that remain un-

“ The prototype design (of the information management system) employs speech recognition, natural language processing, and image recognition technologies. ”

derserved by current technological approaches. Integrating key portions of standard Web browser technology (Netscape Navigator™ and Microsoft Internet Explorer™) with the capabilities of the Informix/Illustra™ ORDBMS, Multimedia Archival Systems has developed a prototype of a client/server system for indexing, cross-referencing, and retrieving data stored in free-form text, audio, and video image formats, along with standard RDBMS data types, to support specific national security decision making,

---

<sup>5</sup> The overall system described has a patent pending and uses existing and pending commercial licenses from Netscape™, Illustra™ Information Technologies/Informix, SUN Microsystems™, and Carnegie Mellon University, issued to Virtual University International, Inc., with sublicenses to MmAS.

<sup>6</sup> MmAS shares in the technological developments of Virtual University International, Inc., which is using similar technologies in its university course content preparation and delivery system.

intelligence assessment, and military command function needs. The prototype design employs speech recognition, natural language processing, and image recognition technologies. Future versions will allow users to create automatic cross-referencing of data records of different types. The proprietary automatic cross-referencing scheme is a kind of correlation coefficient or measure of association for eclectic, nonalphanumeric data and aids in the synthetic, intellectual tasks common to most top-level national security and intelligence policy makers and analysts.

“A standard build-to-specifications at a fixed price approach will almost certainly fail to exploit the latest technology.”

cient or measure of association for eclectic, nonalphanumeric data and aids in the synthetic, intellectual tasks common to most top-level national security and intelligence policy makers and analysts.

Object-oriented programming techniques create extensible systems, both in terms of data types and applications. The Multimedia Archival, Inc., information management system uses COTS technologies and industry standards, and it has the common look and feel of the leading Internet browsers.

The Internet protocols and design standards form the primary client/server architecture and distribution system and it, rather than the computer operating system, functions as the software development platform for the systems outlined here. An “Internet-compliant” strategy tracks the Internet’s rapidly evolving technological advances and, in so doing, maximizes markets and maintains compatibility with the widest array of COTS hardware, communications, software, and information resources. This approach is also referred to as an Intranet system or strategy.

In the context of national security and military intelligence applications, the system described here has two major functions. First is to customize the MmAS client/server information management system to the particular demands of intelligence and national security data. This means creating efficient natural-language and image-recognition-based search engines, user interfaces, and data objects customized for the particular data types used in national security and intelligence systems. The second is to organize data capture operations that will convert raw and analog data—text, images, sensor information—into digital form, which then can be stored and retrieved efficiently using ORDBMS server technology licensed from Informix/Illustra™ (through MmAS).

**IMPLICATIONS FOR THE ACQUISITION PROCESS**

---

## IMPLICATIONS FOR THE ACQUISITION PROCESS

---

If government national security and intelligence organizations are to capture the functional and economic benefits of the rapid revolution in information technology, it must use COTS hardware and software to the greatest extent possible, must use industry standards whenever possible, and must anticipate and follow trends in technology. The task outlined in this paper is a complex systems integration task perhaps best performed by a nongovernmental unit operating under a general, task-defined contract and free to acquire necessary COTS hardware and software, integrating components to meet the general needs of the client. A standard build-to-specifications at a fixed price approach will almost certainly fail to exploit the latest technology.

## **COMMERCIAL APPLICATIONS OF THE UNDERLYING TECHNOLOGIES**

---

Internet-compliant MmAS information management systems have applications in any situation in which the core information for an enterprise or profession does not fit the fixed-record-length, alphanumeric-data requirements of the standard RDBMS model. Commercialization opportunities for the information management system broadly outlined here have been investigated for use in national intelligence, print and broadcast news media, as the course content preparation and delivery system for virtual universities, and for use by health care providers (Electronic Clinical Patient Records/Telemedicine). Current development activity is aimed at the education, distance learning, and corporate training markets.

## **CONCLUSION**

---

Modern information and communications technologies, while offering tantalizing possibilities to military and intelligence professionals, have certainly made

their jobs more difficult and complex. The stove-piping problem inherent in national security and national intelligence organizations is exacerbated by current information and database technologies. Whether a decision maker or analyst is trying to understand context, identify missing pieces of information, or make bets on future scenarios; whether attending to specific real-time needs of a military commander or attempting to identify clear trends and major long-term tendencies of a national system; whether using the methods of a Maigret or a Sherlock Holmes, he or she could use a “technical fix” that helps put together the pieces of diverse, rich information together into a coherent whole.

Can modern information technology help? Recent technological developments, particularly in the software arena, hold considerable promise for enabling national security and intelligence professionals to better cope with the information retrieval and synthesis tasks necessary to perform well in what is often an “overly rich” information environment. The answer is maybe—but not if the sole concern is modifying legacy, relational database systems.

**REFERENCE**

---

Jennings. David L. (1994, December).  
*Multiclassifier fusion of an ultrasonic  
lip reader in automatic speech recogni-  
tion.* Wright-Patterson Air Force Base,  
OH: Air Force Institute of Technology.