Catching bad guys is hard work; doing so while simultaneously protecting civil liberties is doubly hard.



By Jeffrey Pollock

Present capabilities rely on extraordinary people and ordinary IT systems. Approximately 70 to 80 percent of the daily work that goes into finding and catching terrorists, money launderers, Internet hackers, and other criminals is done in front of a computer. Bad guys leave electronic footprints everywhere they go: drivers license records, credit card transactions, telephone logs, Internet usage, airline trips, car rentals, tax records, snapshots of license plates while moving through toll roads, and so on. The trouble is that everybody, not just the bad guys, leaves these same footprints.

business integration journal takeaways

BUSINESS

- Most inefficiency in law enforcement is in the information retrieval process.
- DARPA's Total Information Awareness (TIA) initiative will bring emerging technologies to bear on the issues involved in traditional enterprise integration, yielding improved business integration methods.

TECHNOLOGY

- Technology isn't the central problem with integration today, but emerging technology can be used to alleviate some of the traditional non-technology problems.
- Technologists should better separate the information level concerns from the connection level concerns, improving the looseness and scalability of the overall middleware architecture.
- Emerging tools to assist middleware platforms with more efficient handling of information come in three forms: semantic modeling design tools, registry and thesaurus tools, and mediation, reputation, and inference engines.

This is trouble because there are hundreds of thousands of IT systems throughout the U.S. (and beyond) that store these disparate data records. Getting access to all that data is inefficient, at best, and sometimes impossible. If a super authority were to mandate that all this data be moved to a single physical location, severe risks to civil liberties would exist. The unplanned, chaotic, decentralized nature of the U.S. government's IT infrastructure is both a barrier to security progress and also a protector of freedom and liberty.

Smart people throughout the U.S. government are still trying to sort out a plan to simultaneously help law enforcement and protect civil liberties. Nearly everybody has quickly nixed the big database or supercomputer approach to solve the problem. Instead, various government agencies and local law enforcement agencies are looking at ways to achieve total information awareness, while maintaining significant safeguards against infringement on civil liberties.

Likewise, many other commercially available technologies have also been eliminated. Today, the commercial state-of-the-art simply isn't good enough. The U.S. government is facing an IT problem of unprecedented scope and scale. A problem of this nature is inherently life or death, exists on a political and technological scale, and is daunting to overcome.

Technology Isn't the Problem

When it comes to efficiently sharing intelligence information inside the U.S. government, technology isn't the core problem. A seemingly simple task, information sharing has proved to be a treacherous sinkhole of time and money for our governmental leaders. It's important to note that large business enterprises routinely fail at the same task—upward of 88 percent, according to some analysts.

The reasons for our problem with achieving swift, efficient information sharing are intricate and intertwined, but technology is only a single factor. Others include:

- Operational limitations have perhaps been the most significant barrier to joint efforts on information sharing. Differences in agency objectives have led to differences in processes, equipment, and organizational authority. Along with cultural issues, operational limitations have greatly prevented substantial progress. Thankfully, much of the work done in Congress, the White House, and key executive agencies during the past 18 months has targeted these areas for improvement.
- Budgetary constraints have produced two kinds of hardships. The level of funding has been insufficient and the degree of cross-program coordination for similar efforts has been neglected—particularly as it pertains to the required infrastructure for making various information resources interoperable.
- Cultural differences are frequently the most difficult to spot because they're not technical or procedural problems. Culture is embedded in the way one government agency chooses to operate or the way language differences evolve among organizations. Culture is subtle; it's folklore; it's pride in one's differences. These factors cause difficulties when sharing information.
- Political wrangling leads to fiefdoms of control and dangerous silos of isolated information. People in power are often motivated to accrue additional power. Today, information is power. Where power isn't the prime motivation, many agency leaders worry that ceding operational control of information could compromise their primary mission and contribute to information hoarding.
- Technical incompatibilities at least tend to be black and white in nature. Typical incompatibility issues will boil down to application protocol, programming language, or data definition problems. These are issues that can be fixed, albeit often at significant taxpayer cost.

These challenges can be overcome. As the Markle Task Force Connecting for Security working group has suggested, there are at least 10 major characteristics that the next-generation Homeland Security Information Network should meet. Additionally, the Federal Enterprise Architecture (FEA) Working Group has set baseline technical frameworks intended to unify agency development programs. Other groups, such as the Regional Organized Crime Information Center (ROCIC) have established regional information sharing systems that mostly consist of Internet portal-type applications. These let specialists search for data using keywords within various systems.

As these efforts suggest, there are three primary demands for information sharing:

- To enable visibility into disparate systems, preferably with analytic capabilities to spot trends that develop over time and across systems
- To enable interoperability among these systems directly so local alert systems and agency software applications can operate on data directly from within their own environments
- Include sufficient shielding, security, auditing, and accountability mechanisms to conform with existing and emerging privacy laws. Early steps have been taken to accomplish both of these goals, but mostly by uncoordinated smaller efforts with little cross-mission planning. This focus on tactics, rather than strategy, is due to a dearth of workable technologies that effectively compensate for the non-technical issues at hand.

The full scope and complexity of this problem can be appreciated only when viewed in its entirety. Although federal agencies already have volumes of information that could be shared, no single interagency exchange is technically impossible. Consulting companies would be happy to work on specific intersystem communications. The core challenge for visionaries today is to imagine an infrastructure that can be built, and built upon, in the future—that lets local, state, and federal agencies (as well as various commercial entities) tap into the required information on demand.

Total Information Awareness

Total information awareness, sometimes referred to as information fusion, originated as a battlefield concept. It describes a battlefield commander's need to have access to detailed information regarding all aspects of land, sea, and air forces engaged in a battle. Progress in this area has resulted in the high-tech wizardry found in most of the fighting vehicles and troops stationed in the Middle East today.

This concept of operations has expanded its reach to include aspects of Homeland Security that deal with finding the bad guys here at home. But instead of ordering new equipment for the armed forces, we're faced with architecting a massive IT solution that will impact local, state, federal, and international IT systems deployed across the globe.

The Defense Advanced Research Projects Agency (DARPA) is leading the national research and development program in this area. Initially conceived of

as the Total Information Awareness program, DARPA has bowed to political pressure and misunderstandings regarding the scope of their effort to better define the program as the Terrorism Information Awareness (TIA) Program. Under this program umbrella, there is a mind-boggling array of emerging technologies being examined for use inside an expansive new technology infrastructure. There are three categories of technologies being examined:

- Language translation
- Data search/retrieval
- Decision support tools.

Within each of these categories, there are many different kinds of approaches under consideration, including pattern recognition, artificial intelligence (AI), collaborative networks, semantic architecture, systems interoperability, and so on.

At the detailed level of operations and research, the TIA agenda is taking shape. Typical of DARPA programs, the TIA effort is high-risk, high-reward, and could take many years to achieve results. While the larger community is still defining precisely what TIA is, there are several technologies related to data search/retrieval that we know it isn't:

Single vendor EAI or business-to-business (B2B)-type tools: No commercial entity has ever attempted an information-sharing scheme of this scale. No commercial middleware vendor has built test scenarios for data sharing at this scale. Even if the connectivity portion of the problem were to be accomplished (e.g., hundreds of thousands of systems interconnected), the non-trivial barrier of resolving information incompatibility would remain. Even if we resolved significant portions of the information incompatibility problems, we would have to start all over again to maintain the environment and keep up with changes. A new, non-traditional approach must be attempted to successfully tackle a problem of this scale.

Data warehouse: Initially a popular answer, the central database is the obvious solution. Unfortunately, the technical challenges of distributing, securing, optimizing, designing, maintaining, and implementing a database instance of the required parameters leave too many undesirable technical trade-offs. Moreover, there are political ramifications of providing a single point of attack to potential wrongdoers and possible threats to civil liberties.

Standard XML vocabulary: Standardized vocabularies are hard to agree on, painfully slow to develop, difficult to change, and usually misused and abused. Standards can have a place in a complete solution, but they alone aren't a sufficient answer.

AI for Middleware

While the TIA is a broad vision, the portion of that vision that relates to the EAI industry will redefine the premise of what EAI technologies were built upon. Initially, EAI and B2B approaches evolved from message-oriented-middleware (MOM). As messaging environments became increasingly complicated, new services and features were added to core MOM capabilities. But with the advent of the TIA, the use of the information inside the messages—not messages themselves—is the focal point of the technology development.

AI technologies inside the middleware will increasingly bring focus to the content of the message and its possible usages.

This observation runs contrary to most of the media hype that surrounds our industry today. From Web services to business process management (BPM), we're constantly inundated with vendors telling us that pipes, plumbing, and orchestration of messages is the Holy Grail of integration. But the scale and complexity of the present intelligence community challenge highlights the need to directly address new requirements. These requirements derive from the characteristics and goals of the Markle Foundation Task Force and must succeed at addressing operational, budgetary, cultural, political, and technological problems. Due to the necessity to balance a diverse set of loosely defined concerns and address multiple layers of complexity, what we need from technology in the future is vastly different from what we have today.

A fundamental aspect of a successful solution is to apply the architectural pattern of separating concerns. In the future TIA architecture, the information must be separated from the connectivity technologies as a way to divide and conquer.

Separation of the information from the more traditional technologies makes the architecture flexible, loose, and independent with regard to the core information flowing through the pipes. In this separation, the logical information layer should encompass logical data models, conceptual models, business rules, process models, logical query structures, and all relevant logical-to-physical mapping objects.

From a tools perspective, there are several technologies that will assist both the run-time and design-time functions of this information layer. Recently, we've witnessed several middleware companies taking innovative approaches to solving confounding data exchange problems. Many of these companies apply AI technology directly in these efforts.These newer sets of technologies can be referred to as semantic integration tools. They all focus on the meaning and intent of information to enable more efficient, capable data exchanges in large communities.

A brief survey of emerging tools would consist of three key categories:

- Semantic modeling design tools can be said to add complete, decoupled, robust metadata to source information of any type and format. Many vendors provide some capabilities in this area but there's still considerable fragmentation regarding specific methodologies and approaches. Typical tasks that a modeler or designer would wish to accomplish

Markle Foundation Key Characteristics

1. Empowering local participants: to access, use, and analyze data on the network's edge
2. Funding and coordination: decentralized, but adequately funded, coordinated, and empowered participants
3. Safeguarding civil liberties: guidelines to prevent abuse
4. Eliminating data dead ends: stovepiped data should not lead to a dead end, with no accountable authority
5. Robust system design: minimizing possible points of failure
6. Network analysis and optimization: detecting patterns in the security network may trigger alerts
7. Solid growth and upgrade plan: the use of good architectural patterns ensures maximum longevity
8. Support for legacy infrastructures: a successful network will exploit existing information resources
9. Creating network-aware scenarios: complex models of security scenarios can assist analysts in the field
10. Connected culture: being connected and participating must become second-nature to participants.

— *J.P.*

include: ontology, schema, taxonomy, and semantic map design. Individuals who perform this type of work would include object-oriented analysis and design specialists, data modelers, and other professionals comfortable with abstraction and data relationship types. Recent movement in the standards community to address some of these core design and modeling areas is specifically geared toward the artifact specifications.

- Registry and thesaurus tools act as a link between the design world of models, maps, and taxonomy and the run-time world of software compilers, servers, and engines. Many vendors and standards have emerged to address the registry needs of distributed systems. Likewise, many vendors are currently supplying some type of distributed thesaurus capabilities. However, linking these approaches together to support dynamic, self-configuring, self-healing networks remains in its infancy. Techniques to publish and link synonym and antonym IDs across and within schemas are rapidly maturing. Integration tools developers will be able to start decoupling their proprietary interfaces when a blend of universal description, discovery, and integration (UDDI), lightweight directory access protocol (LDAP), and thesaurus services are available in an open, standard framework.
- Mediation, inference, and reputation engines are a category of tools that provide the run-time capabilities to dynamically generate various execution programs without prior involvement from application programmers or developers. Much can be done to automate time-intensive programming by reading the metadata generated from the semantic modeling and design tools as input to the run-time process. A mediation engine would possess the capabilities to broker query requests, dynamically generate distributed smart queries, and aggregate results from widely disparate sources. Inference engines provide a way to infer the meaning of data that moves through various contexts and assist the query generation process with description logics. Reputation engines provide run-time access to scoring mechanisms for both content quality and service security, enabling a flexible, dynamic blueprint for on-demand interoperability.

Each of these approaches offers tremendous benefits, but when taken together, the whole is greater than the sum of its parts. A TIA platform constructed with these technologies at its core would be a central nervous system of information and succeed in achieving new operational capabilities with a small footprint and lower costs. However promising these emerging technologies may be, we must still address the issues with pipes and plumbing.

Concerning the Plumbing

Plumbing consists of pipes and fittings. A pipe is the vehicle in which contents flow and the fitting is how the pipe connects to joints or other sized pipes. With application integration, the pipes are analogous to network protocols and transport mechanisms. These are the technologies that carry content over geographic distances. Fittings are more analogous to the transport brokers that can be connected to different transport types and can even act as switches between them. So, IT plumbing concerns the physical layers of software, independent from the information intent or meaning.

The physical layer traditionally encompasses:

- Connectivity protocols
- Workflow brokers that execute process and routing models
- Data persistence mechanisms
- Data exchange formats (instance data)
- Physical security rules
- Tools to manage systems participating in exchanges.

Most IT systems relevant to TIA initiatives already exist, have access points, interfaces, and users. This complicates the plumbing concerns because, while AI can enable the information layer to become loosely coupled, the physical layer must usually be tightly coupled due to legacy protocols and transports. For example, an older application serving up international warrants to federal sources may have a back-end interface that supports CORBA over Internet Inter-ORB Protocol (IIOP), but to connect to it and use it, an outside system must bind to its published interfaces and use its objects with help from human developers.

Popular plumbing techniques—such as message queuing, file transfer protocol (FTP), enterprise JavaBeans (EJB), remote method invocation (RMI), Rendezvous, Open Database Connectivity (ODBC), Java database connectivity (JDBC), component object model (COM), distributed component object model (DCOM)—all require a firm handshake and tight coupling between their interfaces. This tight coupling can be manifested in several ways: vendor reliance, programming language reliance, application programming interface (API) contracts, workflow language, and other low-level exchange mechanisms.

Recent progress in Web services has directly targeted some of the challenges with middleware plumbing issues. With Web services, the API contracts are published and registered. In the ultimate vision, new IT systems can discover new pipes and then attempt to connect to their fittings without programmer assistance. This process works well because of widespread agreement that:

- Simple Object Access Protocol (SOAP) is a good piping material.
- Web Services Description Language (WSDL) is a good way to describe its fittings.
- UDDI is a good mechanism for discovering services on the network.

The TIA program is moving toward reliance on a Web services grid as its primary mechanism for pipes and fittings. Within this architecture, techniques for BPM, security, auditing, and logging are all critical to provide a robust infrastructure for the information-based tools to reside on. The Web services grid architecture provides the mechanism to loosely couple the physical aspects of the final TIA solution.

With the semantic integration and Web services components of the architecture in place, a new kind of intelligence information sharing infrastructure is at hand.

Everything (Loosely) Connected

Taken together, the combination of these two emerging categories of integration—one focused on the operational use of shared information and the other focused on transportation of data and connections between applications—can provide powerful new middleware architecture.

The benefits of blending these approaches can be expressed by describing some of the core characteristics of the TIA architecture:

- **Dynamic**: The ability to create new

implementations on-the-fly at run-time without programmer intervention. Critical capabilities include dynamic data transformations, dynamic query generation, dynamic map generation, and dynamic discovery/routing of network services.

- **Real-time**: The ability to avoid the months of development time typically required to establish new information sharing connections. We cannot continue to require that large projects be funded and implemented to link together relevant intelligence applications. For the sake of security, connections need to be made in minutes, not months.
- **Loosely coupled**: The use of indirection and abstraction in the technical process of linking components. There are two kinds of loose coupling that are critical: loosely coupled information/schema and loosely coupled connections/services.
- **Highly flexible**: The ability to connect to and make use of legacy environments that could include legacy applications, legacy middleware, legacy databases, and legacy Web portals. This also implies the ability to use various network topologies such as hub-and-spoke, message bus, publish-subscribe, point-to-point, and grid infrastructures.
- **Secure**: The ability to install and configure network, application, and data security at all levels of architecture. This should include authentication, authorization, application firewalls, encryption, certificates, and auditing and reconciliation.
- **Open**: The quality of the technology that assures minimal dependence on proprietary products or formats, neutrality with regard to business processes and business data formats, and software adapters and connection protocols.
- **Service-oriented**: The always-on nature of a component architecture that awaits invocation by remote, sometimes anonymous client software.
- **Automated and adaptable**: The ability of the middleware platform to learn, adapt, and automatically generate new physical and semantic connections. Key capabilities of this autonomic computing include query generation, data transformation, schema mapping, schema generation, and thesaurus maintenance.

Technically speaking, these combined capabilities will offer the U.S. government a way to connect intelligence systems without requiring massive databases, single vendor solutions, or nightmarish maintenance requirements and expenditures. Here we find a case where new technology can be applied to ease cultural, operational, and political issues.

Catching Bad Guys Gets Easier

Many security analysts say that another event of mass terror on U.S. soil is only a matter of when, not if. Someday in the not-so-distant future, analysts may have near-instantaneous access to information residing in previously disconnected sources that may help deter such an occurrence. Many of the barriers we're confronting today are non-technical, but can be addressed with emerging technologies. As we envision a TIA platform, we must recall the necessity to protect civil liberties and minimize potential misuse of the technology.

Authorities need capabilities to move from reactionary tactics and toward proactive prevention of crime, illegal activity, and terrorism. Being able to recognize patterns in controlled data sets and perform analytics on distributed software owned by various agencies and third parties will quickly become a reality as capabilities continue to mature. It's possible to create an architecture and set of tools that will enable such analysis while still preserving civil liberties. Ultimately, the time and money that go into computer-aided suspect and incident searches should drop as capabilities dramatically improve.

Throughout the myriad of advanced technologies currently within scope of the DARPA Information Awareness program, of which data search and retrieval is only one, there continue to be pressing legal issues. There's a need to collect evidence and documentation of processes anytime enterprise or agency boundaries are crossed. Legal requirements demand clear statements regarding various relationships that can include defining who owns what and what rights they have, in addition to various degrees of trust, liability, service agreements, and privacy.

As we tread the delicate path between protecting citizens, agencies, and enterprises while seeking to improve our national information technology infrastructure, we'll increasingly lean on new technologies to bridge these gaps. Emerging semantic integration tools and approaches, combined with Web service grid architectures, will greatly contribute to an acceptable information awareness solution that benefits all. BIJ

An Information Utility Grid

A popular metaphor in middleware discussions is that of the electrical grid. It's easy to visualize the kind of IT solution that adapts to serving new nodes in the grid no matter what utility supplier provides service. While this metaphor is powerful, it also masks some of the fundamental complexities of deploying middleware in the modern IT environment.

Imagine that your household electrical circuitry needed to alter its power supply to suit any device plugged into it. Applications in IT environments need flexible access to provider services and require that the formats and meaning of the services match precise requirements. What if you needed to plug in both direct current (DC) and alternate current (AC) devices into your wall outlets? What if both the voltage and the amperage of the AC or DC currents were different for each device you plugged in? What if any deviation in the amperage or voltage would cause your device to fail? Finally, what if all these variations and levels of precision needed to be controlled by a commonly shared service utility?

This parallels the kind of complexity we face in designing an efficient information-sharing network. In this example, modern appliances come with adapters that are unique for every device. However, adapters for solid-state electronics can be mass-produced and shipped with each device. That's not the case for business software, where programmers hard-wire and rewire adapters to accommodate their unique and changing environments. So, while the electrical grid analogy is useful, it has serious limitations in helping people understand the layers of complexity involved with middleware solutions.

— J.P.

About the Author



Jeffrey Pollock is an independent business integration advisor. As the former CTO of Modulant, principal engineer with Modem Media, and senior architect with Ernst & Young's Center for Technology Enablement, he has architected, designed, and built application server and middleware solutions for dozens of Fortune 500 *and U.S. government clients. e-Mail: jeff_pollock@yahoo.com.*