

Information Sharing for Homeland Defense:  
Concerns and Suggestions

Kay Hammer  
Evolutionary Technologies International  
October 2002

After the Oklahoma bombing and 9/11, few U.S. citizens question the value of enabling law enforcement, military, and intelligence to share information in order to provide greater protection of U.S. citizens and property. As a result, the general populace appears to be in support of the money and legislation required to support Homeland Defense. However, despite the potential financial benefits afforded the hard-hit technology community, IT professionals must assist the government in understanding the risks associated with such an initiative if our national defense depends upon its successful execution in a timely manner. The purpose of this paper is to enumerate these areas of risk, as well as to suggest several strategies for reducing their impact.

The paper is divided into three major sections. The first discusses the range of challenges that face the information sharing initiative; the second includes a proposal for how one might build a non-invasive delivery architecture that could address many of these risks, and the third addresses how the ETI® product set could either help or serve as a model for portions of such a delivery architecture.

## **I. A catalogue of risks**

There appear to be five major areas of risk:

- Organizational change
- Technical complexity
- Conceptual complexity
- Authorization and access
- Cost and accountability

Understanding the range of issues associated with each of these topics and developing a strategy that balances risk with acceptable results will be critical in determining the best short-term and long-term solutions for delivering acceptable results.

### **Organizational change**

The successes and failures of mergers and acquisitions in the private sector should serve as a cautionary tale as to the potential speed of consolidating government agencies. Corporate cultures and the processes that companies develop to enforce them are at least tempered by the profit motive. Ultimately, commercial entities succeed or fail on the basis of their financial performance and this objective criterion can be used

not only to drive decisions, but to assess their effectiveness. Despite talk about accountability, issues are not so clear in the public sector, where decisions are more likely to require tradeoffs between more complex issues such as security and privacy, control and freedom, etc. Having witnessed the failures resulting from mistakes made in the hasty attempt to integrate previously independent entities, companies like Wells Fargo (in the merger with Norwest) and Sears (with its acquisition of Lands End) have taken a more measured approach, allowing both organizations to work independently for some period of time. Likewise, using Homeland Security as an umbrella organization that governs agencies as diverse as the INS, Coastguard, and Customs will require an equally measured approach, with the end result being a huge increase in costs prior to any efficiencies of scale.

### **Technical complexity<sup>1</sup>**

*NOTE: The following comments regarding technical challenges are far from exhaustive, but suggest the kind of criteria that should be considered in evaluating any technical proposal.*

**Data volume.** Despite the fact that hardware, database and data warehouse vendors are eager to see large data warehouses as the architectural solution enabling information sharing between agencies, the sheer volume of data precludes this strategy. An example of such a massive governmental data integration project would be Centrelink. In 1997, Centrelink was established to consolidate Australian citizens' access to government services. Data from over 40 systems owned by 14 government agencies was integrated into a data warehouse that is currently growing at a rate of 2.5 terabytes a week. One of the most complex sources consisted of a file with over 1.25 billion records, each of which contained over 3,000 fields.

Australia has approximately 19.4 million citizens (July 2001 estimate); the U.S. has approximately 278 million citizens (July 2001 estimate). Extrapolating from the Centrelink statistics, a similar data warehouse for benefits information from federal agencies providing human services in the United States – for example, Social Security, Medicare and the Department of Veterans Affairs – would exceed the computing capacity of even the most powerful commercially available database systems.

**Unstructured data.** Some government agencies, such as those that provide human services, store information comprising fairly straightforward transactional data, such as births, deaths and applications for benefits. In this domain, it is relatively easy to define a meaningful event. However, the types of data being captured by intelligence agencies are far more complicated. In the intelligence field, vast amounts of random data are

---

<sup>1</sup> A number of the points made in this section were previously made by the same author in "Another View: Can government integrate all its databases?" in the June 03, 2002 issue of *Government Computer News*.

gathered with the hope that some pattern may emerge as meaningful. Much of that data is unstructured and comes from proprietary systems. Necessarily, it must be accessed through proprietary interfaces or reporting systems.

For years, the computer industry has been developing technology for capturing and analyzing unstructured data, such as email text, satellite images and speech. However, the integration of this technology with commercial database management systems is still in its infancy. Commercial database management systems support the storage of unstructured data, however, the data needs to be manipulated, i.e., analyzed for patterns. There are a number of vendors that are beginning to offer XML database management systems. (XML is a standard mark-up language for representing documents.) These XML database products use rules provided by the organization to index and store XML documents in order to provide rapid access to particular documents through their search engine. All of the database vendors already have or will be adding these capabilities as part of their core product offering.

It is important to note that while these products will be helpful, they are far from a complete solution, as the effectiveness of the search capabilities depends, in part, on the rules used to create the indices. Furthermore, specifying these rules requires expertise on the part of the programmers deploying the product as well as those who understand the domain – in short, a task similar to training the “expert systems” created by Artificial Intelligence vendors in the 1980s. Finally, text is only one form of unstructured data captured in intelligence systems – audio, graphic, and multimedia information are also gathered and will require similar “expert” storage and retrieval capabilities. It may take several years for these multi-media databases to mature sufficiently to provide both the performance and sensitivity required to execute complex queries. In the meantime, it is likely that most agencies already have proprietary storage and search programs for this kind of information. In the short-term then, it would be advisable to devise a means of invoking these proprietary applications via some multi-media query mechanism.

**No single solution.** One of the most frequent factors leading to the failure of new technology in the marketplace is that it is designed without considering what it will take to effectively deploy this technology in the target customer’s environment. For example, a highly touted, new technology, Web Services, promises to provide a much more dynamic solution to the problem of realtime application integration than the earlier technologies. However, Web Services are not yet sufficiently deployed in the marketplace for the pitfalls to be well understood. For example, consider the case with middleware, which promised to solve the enterprise application integration (EAI) problem. In 2001 the industry analysts Forrester Research reported that companies they interviewed indicated that for every dollar they spent on middleware, they were spending between \$10 and \$15 in implementation. Given the importance of delivering a successful result within as short a period of time as possible, it is critical that any strategy recognize the short-term pitfalls from over-engineering the solution.

**Query Prioritization.** Regardless of the delivery architecture adopted, one of the large technical unknowns is the burden that inter-agency requests will make on the systems in

question and how to prioritize requests. The importance of this aspect in the design of data warehouses in the commercial world is well-understood and accounts for the success of companies like Business Objects and Cognos that provide administrators with a way of tailoring canned reports for management that can be accessed from intuitive graphical interfaces. Because a badly structured query can cause relational

database management systems to seriously impact other users by slowing down response times, few companies allow managers to create their own queries and reports against data warehouses. These “canned reports” both provide greater ease of use – and minimize the impact of query execution on the performance of the system.

### **Conceptual complexity**

***A common data model.*** Given the challenges outlined above, the best means of inter-agency data integration may be some form of virtual warehouse that allows an agency to formally characterize the kind of information it is seeking without divulging the underlying structure and content that would reveal how the information was obtained. In a virtual warehouse, users would have a query environment that allows them to ask interesting questions against an abstract data model – that is, a logical view of the data – rather than against an actual database containing data values. The virtual warehouse would translate any query issued against the abstract data model into the appropriate set of actions to extract and consolidate the data from the agency systems in order to satisfy the request (providing that the individual issuing the query had the appropriate authorization). However, there is considerable risk associated with such a data model if it seeks to represent a consolidated view of all the (shareable) data elements used by all the applications in every agency because of the conceptual complexity of deriving such a common data model. In fact, this type of task can be so time-consuming that commercial organizations often purchase an industry-specific data model (e.g., for financial services or healthcare) from vendors like IBM to use as the target schema in their enterprise data warehouses. Given the range of the information to be shared – from the location of ships to parking tickets – a less risky approach might be that of an intelligent card catalogue where individuals would first choose the type of data source they want to query (e.g., Coast Guard or local and/or state governments) prior to requesting the desired data. This would eliminate the need to capture the potentially complex mapping between the common data model and the data layouts actually used by the systems in question.

***Importance of metadata for change management and auditability.*** One could envision a virtual warehouse that consists of an intelligent card catalogue that not only describes the different types of information available from various sources, but how that information relates to other data sources in the catalogue. Some graphical query interface would assist parties in generating their requests, which would then be translated into remote queries and reassembled into a consolidated report once the information had been acquired. It is important that the architecture adopted be metadata-driven so that it is easy to determine the impact of any change to one or more of the data sources since timely impact analysis will be required to keep the query

interface “in sync” with the systems to which it interfaces. Likewise, using this metadata (that is, data describing data or processes) to keep an audit trail of what information has been provided to which individuals and/or organizations could be important to recognizing unauthorized access to the system.

**Authorization and access.** Clearly, any proposed solution must support strong security against unauthorized access, a task that is significantly more complex than the usual functionality provided by database management systems. There could be several approaches to this ranging from encryption to document management systems that provide secure delivery and can restrict the ability to copy or the number of times it can be read. Because of the potential damage that could be caused by unauthorized access to secure data, addressing these concerns will significantly affect what is acceptable as an architecture for the information sharing initiative.

**Cost and accountability.** Of necessity, the IT budget for Homeland Security will be huge, as will the complexity of the project scope. Because most of the work – including the design – will be outsourced, it will be critical to have a rigorous review process to ensure that the strategy taken with respect to infrastructure is one that fulfills the various users’ requirements as quickly and cost-effectively as possible. It will be critical that all parties – those responsible for technical oversight and public opinion – provide a truthful and consistent estimation to the American public, whose faith in the economy and business leadership has already been seriously shaken. In order to achieve this goal, it will be important for those responsible for the technical architecture and delivery to articulate their thought processes and obtain sufficient legislative and executive buy-in -- and that this process receive sufficient publicity – to minimize the chances of public “second guessing” for political gain.

## II. Strategies for reducing risk

**Defining classes of users and their various needs.** Unfortunately, like too many CEOs, the government proponents of the need for information sharing are not technical, and their vision of what it will take to deliver results may not be realistic with what can be delivered in a timely fashion. An over-reliance on technology at the expense of facing the realities of what it will take to deliver on that promise – cultural and technical – could result in the kinds of failures encountered in the commercial sector that contributed to the current economic downturn. (Recall Jacques Nasser’s interview in *Automotive News* last December, where he lists an over-reliance on e-commerce as one of his top five mistakes at Ford.)

Key to avoiding this pitfall is a clear understanding of what *would be good enough* for various users in the law enforcement and intelligence agencies. Some information must be available in realtime or the moment is lost – for example, the names of potential terrorists at car rental agencies or transportation terminals. Access to other information – dealing with the more speculative side of intelligence work, such as emails, unusual fertilizer purchases, etc., could probably have a latency of anywhere from 2 hours to 2 days after the initial query was issued before a result was “too late” to help.

In short – as within industry – our first efforts should be spent on understanding what would be a sufficient improvement with respect to what the various user communities have today in order to reduce our risk of another surprise attack. Then, armed with these minimal requirements, the technology community can design the best short term solution – in terms of time, risk and cost – to meet or exceed these needs as the first version of delivering on the information sharing target, while at the same time designing the next generation – more optimal – solution.

In the area of scientific discovery, Occam's Razor has been interpreted to mean that when there are two competing theories which make exactly the same predictions, the one that is simpler is the better. Perhaps in the case of information sharing for homeland security, the interpretation would be better cast to mean that when two solutions deliver (acceptable) results, the one that requires the least change is the most likely to succeed.

**Using a non-invasive architecture for initial approach.** The following section suggests one possible approach for an architecture that might be “good enough” for the initial delivery. This is based on the principle that given the complexity of the task, the solution that is most likely to succeed in the initial phase is the one that requires the least change. As a result, this proposal was designed to meet the following goals:

**Minimize**

- the investment in new software and hardware
- impact to organizations
- development resources required

*Strategy:* Use a product like the ETI product set that provides the ability to cost-effectively generate native queries to the systems already in use at the various agencies. Key to this is a data-driven architecture to the code generator that can be extended to address 1) proprietary systems and 2) application-specific business rules and search criteria, as efficient response requires that the queries be generated and executed without human intervention.

**Maximize**

- ready access and ease of use
- visibility into who's asking for what

*Strategy:* Using a Web-enabled user interface would result in a delivery system that would provide easy access to the full range of users envisioned, ranging from intelligence agents to local sheriffs or airline agents. For security purposes, the ready access must be offset by a means of tracking, monitoring, and (perhaps) intercepting particular requests.

**Enforce**

- security of information
- policy

*Strategy:* Unauthorized access must be prevented at all costs. Two means of achieving this are by having the user interface “tuned” to the users’ access rights so that users can only “see” the data sources that they are authorized to query. Maintaining a data warehouse where information requests are automatically monitored and flagged prior to fulfillment would allow human intervention in the case that they seem suspicious. Finally, by using native access, the security systems in place at the various agencies provide another level of protection. Document servers accessible through the Web provide a further level of protection in terms of enforcing constraints like how many times a document can be viewed, whether it can be copied, downloaded, etc.

In terms of implementation, policy will be realized in terms of defining which classes of users can access which classes of information. Other requirements on the delivery system might entail such things as how much data can be requested or the priority of processing the request.

### ***Enable***

- efficient change management
- technological innovation.

*Strategy:* A centralized repository should be used to capture the following: the rules/templates for driving the user interface and generating queries to systems, the access rights to data sources by class of user, and a history of all requests made and fulfilled by the delivery system (The latter might be offloaded to a data warehouse for analyzing to detect patterns indicating inappropriate usage.) At the heart of any data-driven architecture, the repository is key to change management and flexibility, providing four major capabilities:

- The creation of temporary data stores for use by investigative teams. By keeping query results within the delivery system and providing access via a Web-accessible document server, cached information for case management – which can take from weeks to years – would have the same level of access protection as the original source systems, reducing the risk of classified information falling into the wrong hands.
- If policies regarding data access change, a GUI that is driven off the repository will automatically reflect the differences. The history of query requests could also be viewed and monitored to determine which queries should be terminated/invalidated.
- Similarly, the repository could be updated with information that could allow the GUI to guide the user in searching proprietary data sources. For example, consider the case where an agency has an application that can process and index a large volume of textual documents. Users cannot ask for the documents they would be interested in without understanding the categories used in the indexing process, and these categories may differ with different classes of documents (e.g., emails versus articles versus

technical documentation). The ability to add a “dataset” and a set of categories or rules for querying it that would allow the GUI to guide the user in search requests would be critical to providing authorized users with an efficient means of getting what they want.

- Finally, these same capabilities would allow agencies to upgrade their systems with little downtime for the delivery system.

### ***One possible configuration***

Figure 1 illustrates the major functional components of a non-invasive delivery architecture, based on the assumption that a Web-based architecture (using both Internet and intra-nets) is probably the best platform for access and delivery. Such a system could be based on a single physical platform or not. But in order to meet the goal of minimizing the impact on existing systems and organizations, it should require little or no software to be installed on the existing data sources.

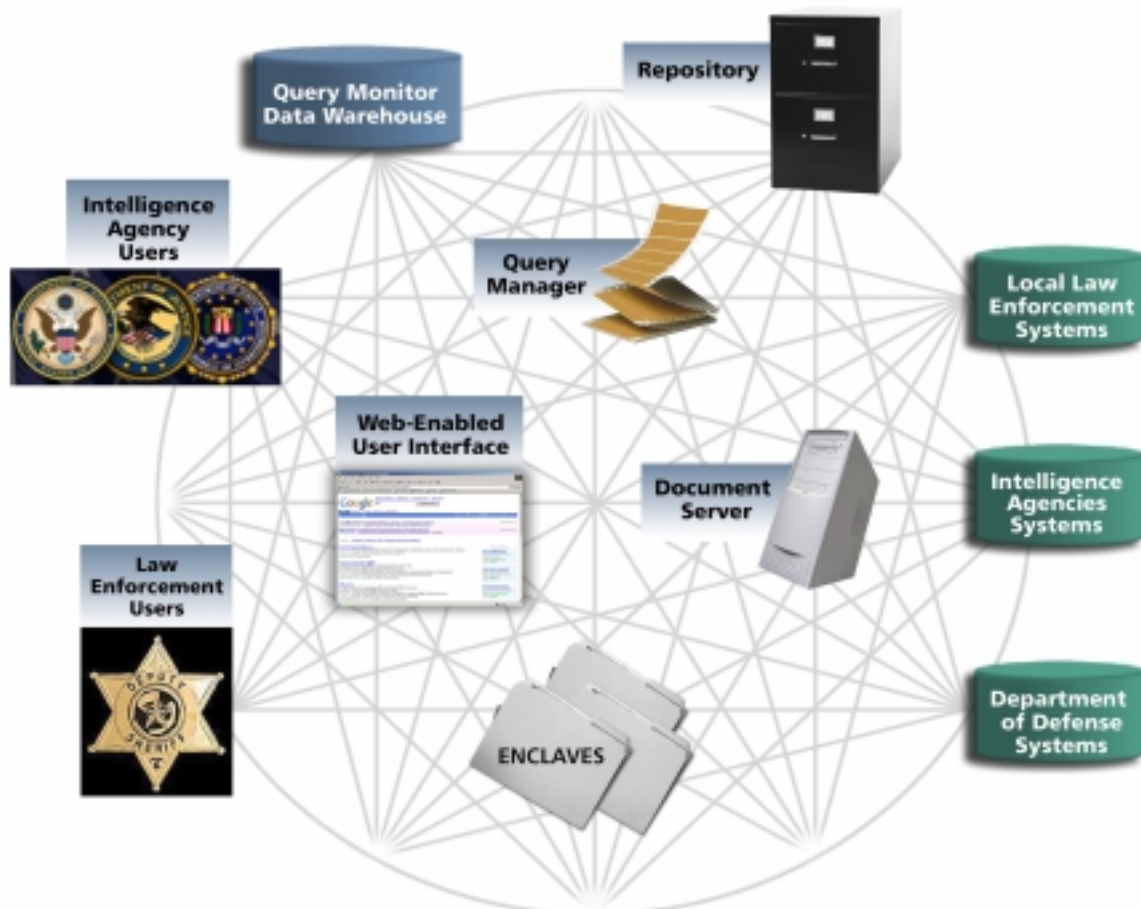


Figure 1

The following sections discuss requirements for the various functional components.

### ***Administration Tools***

Although not highlighted in the figure, a set of administrative tools is required in order to support the creation and maintenance of this type of data-driven delivery system. These tools fall into two classes:

- A set of editors (e.g., specialized tools like a spreadsheet) that would allow the administrators for the various agencies to create the rules, templates, and scripts required to define a data source to the repository, e.g., the schema, query templates, and search criteria/rules.
- A set of administrative tools for maintaining the repository and temporary data sources created as the result of the execution of certain queries.

### ***Repository***

- Must be able to be queried (for the purpose of impact analysis)
- Must be able to version the definitions of systems/queries for purposes of audit or access to historical data. For example, if an agency implements a new system, it may choose not to migrate historical data to this system. Should the need arise to search the historical data, the delivery system could revert to the appropriate access information to drive both the GUI and the Query Manager.

### ***Graphical User Interface (GUI)***

- Must support the appropriate access authorization to restrict the user's view.
- Must provide users with an easy to use means of understanding the data sources at their disposal.
- Must have a means of guiding the user through the specification of search conditions to insure that the search is both semantically viable and correct by construction.
- Must provide users with a means of requesting the timeframe and type of report they need – e.g., realtime authorization of identity, a temporary status report, the creation of a data cache.

### ***Query Manager***

- Must map the user specifications to the GUI into the native queries required by the various sources and launch the queries to the appropriate systems, as well as the appropriate instructions to the document server and database server.
- Must update the query warehouse with the requests being serviced. If there is a desire to put in place another means of validating users/queries in order to prevent misuse of the system, the query manager might wait—on certain types of queries—to receive approval before launching the queries to the source systems.

### ***Query Warehouse***

- The query warehouse would allow the creation of a data mining application to look for patterns that would suggest misuse and then do such things as halt the execution and launch an alert for human intervention.

### ***Document server***

- Based on criteria stored in the warehouse with respect to the sensitivity of the data – with respect to classification or timeliness (for example, certain geospatial information could become obsolete within a relatively short period of time), the document server must enforce such things as whether a user could download or copy a report or document, how long the query results would be maintained, impose another level of user identification before “delivery,” etc.
- Must be able to interface to any temporary data caches created for enclaves for the purpose of case management.

### ***Database server***

- The database server is used for temporary data stores that are created, updated, and/or deleted by commands generated by the Query Manager.

## **III. Parallels in the architecture of the ETI Product Set**

The ETI product set offers the following capabilities:

- Web-based access (via SOAP and WSDL) to an intuitive data-driven GUI.
- A patent-pending means of using a combination of grammar rules and context information to dynamically create menu-driven interfaces that allow users to specify arbitrarily complex business rules or test and transformation logic.
- A means of using user specifications, schema and network information, and templates to create native interfaces and the necessary command language or scripts required to execute the interfaces.
- A repository that contains the data necessary to support the above activities supports versioning, change notification, and impact analysis.
- A set of editors for the creation of the grammars and templates required for the above activities.
- A set of editors used to define user groups, access rights, etc.

Many of the components listed above are suitable, as is, to be used in the type of architecture described in the previous section. Others would need enhancement. For example, ETI's access model is currently not rich enough to support specifying the types of information required to drive the documents server – e.g., with respect to whether a report can be downloaded or should be destroyed after viewing, etc.

October 2002

Should such components be of use in fulfilling the requirements for information sharing for Homeland Security, ETI would – with the appropriate NDA – be glad to discuss its architecture in greater detail with the appropriate parties.