*Rapid Assessment Process (RAP) Report*
*# 04 - 04*

---

# *Operational Net Assessment Enabling Technology: Data Integration*

**Project Alpha**

**Joint Futures Laboratory**

**Project Alpha Team Members:**

Kevin Brandt, Paul Fernan, Christian Grant and Russell Richards

**December 2004**

**Data Integration: An ONA Enabling Technology**

This report, RAP #04-04, was prepared by Project Alpha, Concept Exploration Department, Joint Experimentation Directorate (J9), U.S. Joint Forces Command.

Team Members: Kevin Brandt, Paul Fernan, Christian Grant, Russell Richards

Note: This RAP report is intended to articulate ideas that might enable a future Joint operational concept or provide a new capability for Joint forces. The intent of the RAP report is to give visibility to the ideas, to generate discussion, and to stimulate action in bringing the ideas to fruition.

The views expressed in this report are those of the Project Alpha Team and do not necessarily reflect the position of U.S. Joint Forces Command or the Department of Defense.

**Approved by**: _____

Dr. Russell Richards
Director, Project Alpha
Russ.richards@je.jfcom.mil
757-203-3651

**Approved by**: _____

Shane Deichman
Department Head, Capabilities Department
Shane.Deichman@je.jfcom.mil
757-203-3646

# Data Integration: An ONA Enabling Technology

## *Table of Contents*



## *Figures*



## *Tables*

# Data Integration: An ONA Enabling Technology

## *Executive Summary*

Operational Net Assessment (ONA) is one of the concepts being prototyped by the Joint Futures Laboratory at the U.S. Joint Forces Command (JFCOM). The gist of the ONA concept is to develop a deep knowledge and understanding of a potential or actual adversary by constructing a system-of-systems view of that adversary. Development of an ONA requires vast amounts of data and access to distributed, heterogeneous data sources – many of which are owned by other government organizations outside the DoD, non-government organizations, multinational partners, or private industry. Some are even open sources. Many of these had previously rarely (if ever) been accessed by military analysts.

The challenge to the analyst is to access and integrate data from the disparate data sources to produce a coherent picture that informs the ONA. Easy retrieval of integrated data from multiple distributed, independent, heterogeneous data sources is critical for many other military applications as well. Data integration requires combining and matching information in different sources, and resolving a variety of conflicts. With the number of data sources growing very rapidly, data integration is becoming more important every day. Admiral Giambastiani, Commander of JFCOM, has recognized data integration as one of the command's top priorities. The logistics community is devoting significant resources to address data integration, and it is a major focus of the Horizontal Fusion project. Net-centric warfighting and JFCOM's EBO, ONA and Collaborative Information Environment (CIE) concepts will all depend on our ability to develop good solutions to the data integration problem.

This paper reviews the three traditional approaches to data integration – (1) data warehousing, (2) federated databases and (3) enterprise search. In addition, we considered a promising middle-ware technology, called the External, Index and Query (EIQ) Server. The EIQ Server approach appears to have advantages over the traditional approaches, particularly for the situations expected when conducting an ONA. The paper reports on a data integration pilot study that was conducted as a partnership between Project Alpha and the ONA prototyping team. The pilot study demonstrated the ability of the EIQ Server to process a variety of simple and complex queries against five data sources recommended by the ONA team. It also demonstrated the ability of the EIQ Server to provide near real-time index updates.

The Defense Advanced Research Agency (DARPA), the Defense Information Systems Agency (DISA), and industry promise long-range solutions to data integration that posit transparent data migration. These long-range solutions presently center around the use of extensible markup language (XML) and the development of metadata repositories for several different communities of interest. However, we believe that the EIQ Server offers a very promising near- to mid-term solution to the data integration problem. Furthermore, we believe that it warrants further experimentation as an enabling technology to support the ONA process and similar concepts requiring the integration of large amounts of disparate data from multiple sources.

# Data Integration: An ONA Enabling Technology

## *The Problem*

Operational Net Assessment (ONA) is one of the concepts being prototyped by the Joint Futures Laboratory at the U.S. Joint Forces Command (JFCOM). The gist of the ONA concept is to develop a deep knowledge and understanding of a potential or actual adversary by constructing a system-of-systems view of that adversary. The system-of-systems view includes the following subsystems: political, military, economic, social, information and infrastructure (PMESII). This deep knowledge and understanding is a key enabler of effects-based operations (EBO) which include a variety of actions against the PMESII systems. The action space includes diplomatic, information, military and economic (DIME) actions. An ONA attempts to understand each of the PMESII systems, the relationships and interactions among those systems, and the effects that actions across the DIME domains will have on those systems and relationships. The ONA permits the identification of critical nodes and actions and the resources needed to influence the adversary and shape a campaign. The ONA attempts to understand not only the first order effects, but also higher-order effects and lagged effects.

Clearly, the task of developing an ONA is daunting. The difficulty lies in the vast amounts of data that are required to perform the assessment and the fact that much of the data is not owned by the analysts. An ONA requires access to many different databases that heretofore had rarely been accessed by military analysts. No longer is it possible to obtain all of the data needed by relying only on military-owned databases. Instead, the databases that need to be accessed will frequently be owned by other federal government organizations (e.g., the State Department, Federal Bureau of Investigation, Central Intelligence Agency, Commerce Department, Federal Emergency Management Agency), by non-government organizations (Red Cross, Peace Corps, Doctors Without Borders), by our multinational partners, by our state and local governments, by private industry, or even open source databases. The challenge is to integrate the data from disparate data sources to produce a coherent picture that informs the ONA analyst.

Easy retrieval of integrated data from multiple distributed, independent, heterogeneous data sources is critical for the ONA analyst and for many other applications. Data integration requires combining and matching information in different sources, and resolving a variety of conflicts. With the number of data sources growing very rapidly, data integration is becoming more important every day. Admiral Giambastiani, Commander of JFCOM, has recognized data integration as one of the command's top priorities. The logistics community is devoting significant resources to address data integration, and it is a major focus of the Horizontal Fusion project. Marian Cherry, the Horizontal Fusion Portfolio Manager in DoD's Network and Information Integration Office, characterized horizontal fusion as *"… the ability to integrate data from disparate sources for rapid and effective decision-making to make U.S. forces less vulnerable and more lethal*.[1]"

Net-centric warfighting and JFCOM's EBO, ONA and Collaborative Information

[1] American Forces Information Service News Article by Rudi Williams, September 26, 2003.

Environment (CIE) concepts will depend on our ability to develop good solutions to the data integration problem.

## *Information Interoperability*

Many organizations are looking for faster, easier ways of sharing information via computer systems. The goal is to make information available that data sources have and are willing to export. There are two main types of information interoperability:

- *Exchange*, in which a producer provides information to a consumer, and the information is transformed to suit the consumer's needs.
- *Integration*, in which in addition to be transformed, information from multiple sources is also correlated and fused so that the consumer sees a single, coherent view rather than all the systems' opinions.

Seligman and Rosenthal[2] advocate a framework for information interoperability. The framework addresses four problems. They are:

1. Overcome geographic distribution and infrastructure heterogeneity.
2. Match semantically compatible attributes.
3. Mediate between diverse representations.
4. Merge instances from multiple sources.

***Distributed heterogeneous data sources*** range from conventional databases on a local area or wide area network or intranet to web-based sources across the Internet. Figure 1 depicts the variety of data sources that can face the ONA analyst.

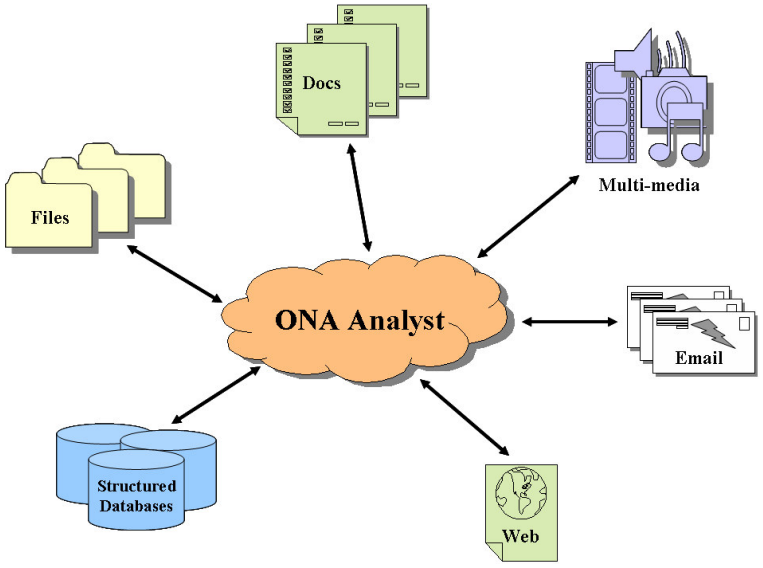


***Figure 1. Heterogeneous Distributed Data Sources***

The databases can be ***heterogeneous*** in many other ways. For example, the various data sources can:

[2] Seligman, Len and Arnon Rosenthal. "A Framework for Information Interoperability," The EDGE: MITRE's Advanced Technology Newsletter, Summer 2004, Volume 8, Number 1, pp. 3-4.

- Run on different hardware
- Use different network protocols
- Have different software to manage the data stores
- Have different query languages and different query capabilities
- Have different data models (different names for the same fields, different units of measurement, etc.)
- Be relational or non-relational (object oriented, flat files)
- Be structured (database files), non-structured (documents and email) or semi-structured (XML, spreadsheets)
- Have different classification levels (unclassified, confidential, secret, etc.)
- Range from conventional databases on the LAN or intranet to web-based sources across the Internet.
- Have different update rates

Another challenge results from the ***semantic incompatibility*** of data attributes in different databases. Some independently developed information systems use the same terms for the same data entities, but many do not. Sometimes the differences are quite subtle. If users combine results across systems without understanding the subtleties, the resulting data is unlikely to satisfy the needs of the application. Yet another integration problem results from the need to ***reconcile different representations of the same data entities***. For example, one database might represent altitude in meters while another measures it in feet. These types of differences are frequently addressed by developing translators (adaptors) across systems.

The ***DoD Net-Centric Data Strategy*** published in May 2003 is viewed as a key enabler of the Department's Transformation by establishing the foundation for managing the Department's data in a net-centric environment. The key attributes of the Strategy include:[3]

- *Ensuring data are visible, available, and usable when needed and where needed to accelerate decision-making*
- *"Tagging" of all data (intelligence, non-intelligence, raw, and processed) with metadata to enable discovery of data by users*
- *Posting of all data to shared spaces to provide access to all users except when limited by security, policy or regulations*
- *Advancing the Department from defining interoperability through point-to-point interfaces to enabling the "many-to-many" exchanges typical of a net-centric data environment*

*The strategy also introduces management of data within communities of interest (COIs) rather than standardizing data elements across the Department. COI is the inclusive term used to describe collaborative groups of users who must exchange information in pursuit of their shared goals, interests, missions, or business processes and who therefore must have shared vocabulary for the information they exchange.*

[3] John P.Stenbit (CIO, Department of Defense) memorandum on "DoD Net-Centric Data Strategy," May 9, 2003.

The Defense Advanced Research Projects Agency (DARPA), the Defense Information Systems Agency (DISA), and industry promise long-range solutions to information interoperability that posit transparent data migration. These technologies include both ***Extensible Markup Language (XML)*** applications and the ***ontological web language (OWL)*** extensions being developed, defined and refined by DARPA and others across the DoD and industry. XML can be used by many kinds of systems that share information via messages – from intelligence summaries to air tasking orders to supply requisitions. XML allows users to define, validate, transmit and interpret many types of applications. With XML, users structure their data using a standardized set of rules, called a schema. Messages in XML are more consistent and precise, thereby enabling quick access and understanding by both humans and machines. The schema are incorporated into the DoD XML Registry which will make metadata available in a uniform, well supported syntax, increasing the opportunities for reuse in DoD command and control systems[4].

The final problem addressed in the framework for information interoperability is that of ***merging instances from multiple sources***. This is usually done through data correlation and data-value reconciliation (***data fusion***). Data correlation determines if two objects, usually from different data sources, refer to the same real-world object. Even when two entries from different data sources refer to the same object, the sources can disagree about particular facts. This requires data-value reconciliation to determine what values the search should return to the application. Unfortunately, this capability requires detailed application knowledge and does not lend itself to automated solution.

## *Data Integration*

Given the wide variety of data sources available, as represented in Figure 1, some software solution is required to efficiently integrate the information for it to be of value to the user.

A ***data integration system*** is an ***automated*** method for querying across multiple heterogeneous databases in a uniform way. In essence, a mediated schema is created to represent a particular application domain and data sources are mapped as views over the mediated schema. The user asks a query over the mediated schema and the data integration system reformulates the query into a query over the data sources and executes it. The system then intelligently processes the query, reading data across the network.

A data integration solution should be transparent to the user. It should mask from the user the differences, idiosyncracies, and implementation of the underlying data sources. Ideally, it makes the set of different data sources look like a single data source to the user. The user should not need to be aware of where the data is stored, what language or programming interface is supported by the data source, if structured query language (SQL) is used, what dialect of SQL the source supports, how the data is physically stored,

[4] Miller, Robert W., Mary Ann Malloy and Ed Masek. "Formatted Messaging Modernization Exploits XML Technologies," The EDGE: MITRE's Advanced Technology Newsletter, Summer 2004, Volume 8, Number 1, pp. 18-19.

or whether it is partitioned or replicated, or what networking protocols are used. The user should see a single uniform interface.

Disparate data and information integration and sharing is all about control over access, indexing, query processing, and result-set data. Various approaches to data integration have been developed. The conventional approaches can be grouped into three broad categories: (1) data warehousing, (2) federated databases and (3) enterprise search.

***Data Warehouse Approach***

Under the ***data warehouse*** approach, administrators define a global schema (template) for the shared data. They provide the logic to reconcile data and pump it into a centralized repository (the data warehouse). This requires an ***extract-transform-load (ETL)*** process that extracts the data from the identified original databases, transforms disparate formats into a single format and loads the results into the data warehouse. The user queries are then directed to the data warehouse and results are returned. Typically, the warehouse is read only, with updates made directly on the source systems. In some cases, however, data marts give individual communities their own subsets of the global data.

The data warehousing approach assumes that the databases selected in advance for integration will allow the ETL process.

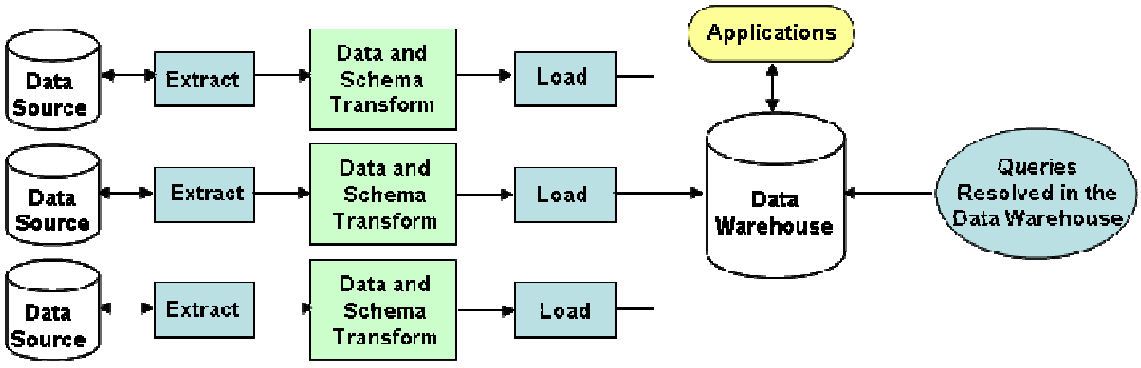


***Figure 2. Data Warehouse Approach to Data Integration***

Depending on the update rates of the databases and the timeliness of the data, data latency can be an issue for the data warehouse approach. An hourly update rate might be suitable for a package delivery tracker but unacceptable for a weapon system needing targeting data.

*Table 1. Advantages and Disadvantages of the Data Warehousing Approach*

| *Data Warehousing Approach* | |
|---|---|
| ***Advantages*** | ***Disadvantages*** |
| Efficient queries on a "single" data source | ETL needed; ETL accounts for up to 80% of the work |
| Only relevant data are stored | Generally requires moving significantly more data than required |
| Supports limited ad hoc queries | Does not allow detailed drill down |
| Optimized for moving and integrating large data sets | Requires significant additional storage |
| | Generally is not real time |
| | Handing the integration of multiple data sources is difficult |

### *Federated Database Approach*

The ***federated database*** approach is really a virtual data warehouse. However, instead of populating the global schema, as in the data warehousing approach, the source systems retain the physical data and a middleware layer (adaptors) translates all requests to run against the source systems. The adaptors assume control over the query process. However, query performance and index types are mainly under control of the data sources. Adapters are usually custom-designed for every data source and involve a "one size fits all" front-end query data schema. Access controls can be put in place. The federated database approach is depicted in Figure 3. The adaptors convert general queries to the correct syntax for the individual databases, individual database queries are executed, and individual database results are processed and presented in a universal format.

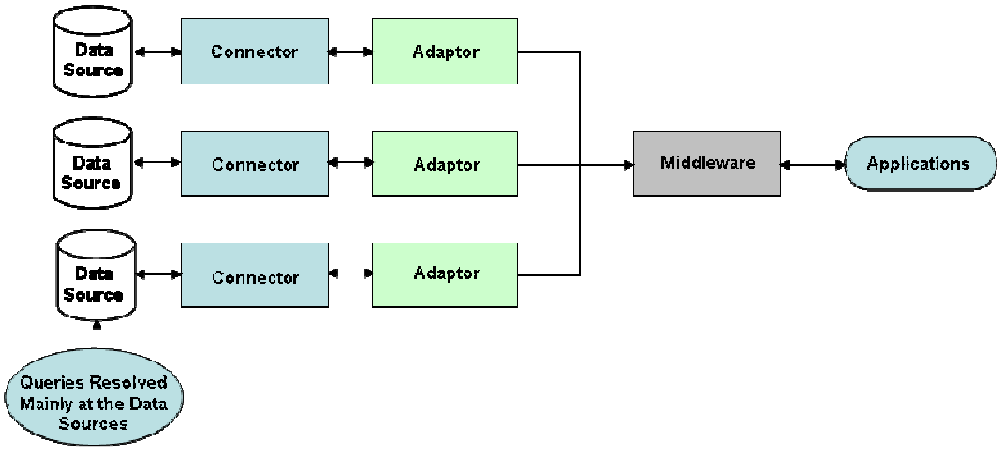


*Figure 3. Federated Database Approach to Data Integration*

Typically a data source has existing applications and users. It is important, therefore, that the operation of the source is not affected when it is brought into a federation. Existing applications should run unchanged, data should be neither moved nor modified, interfaces should remain the same and no additional software should be installed on the machine that hosts the data sources. The way the data source processes requests for data should not be affected by the execution of global queries against the federated system, though those global queries may touch many different data sources. Likewise, there should be no impact on the consistency of the local system when a data source enters or leaves a federation. Communication with the data source is via a client-server architecture using the source's normal client.

Table 2 summarizes the advantages and disadvantages of the federated database approach to data integration.

***Table 2. Advantages and Disadvantages of the Federated Databases Approach***

| ***Federated Databases Approach*** | |
|---|---|
| ***Advantages*** | ***Disadvantages*** |
| No additional storage required | Requires the development of an adaptor for each database in the federation |
| Known, established systems | Exact queries are needed |
| Well suited to queries that require extremely fresh data | Complete understanding of the individual databases is required |
| Enables moving only the data that the user is authorized to see | Generally, federated database adapters are built to accommodate a fixed set of queries, so they tend not to accommodate "unanticipated queries" and "odd combinations" of data |
| Provides near real-time integration | May be difficult to execute complex queries across multiple different database schemas and fields |
| | Queries put loads on source systems at unpredictable moments, thus complicating load balancing and capacity planning |
| | Slower than queries placed directly against a physical database |
| | Significant overhead required to connect to multiple databases, execute queries and receive data from each, consolidate data into a single result set and pass a result to the querying application |

### *Enterprise Search*

An ***enterprise search*** creates an index of the different data sources, searches the indices, and provides metadata and the source document link as the results of user queries. This is the approach that Web search engines like Google or Yahoo use. An enterprise search is very useful for determining the amount of information available on a subject. Further refinements to initial results can assist in shaping an overwhelming amount of data into a manageable amount.

Table 3 summarizes the advantages and disadvantages of the enterprise search approach to data integration.

***Table 3. Advantages and Disadvantages of the Enterprise Systems Approach***

| ***Enterprise Search Method*** | |
|---|---|
| ***Advantages*** | ***Disadvantages*** |
| Very fast | Additional storage for indexes |
| Very comprehensive | Source data needs processing |
| Accommodates multiple file formats | Data can be stale – depends on refresh rate |
| Requires little knowledge of content and structure | Cannot accommodate complex database queries and probably not range queries |
| Can accommodate very large volumes | Unstructured |
| Can accommodate very complex and ad hoc Boolean-type searches | |

### *WhamTech External Index and Query (EIQ) Server Approach*

WhamTech, Inc. (http://www.whamtech.com) has developed an alternative approach to data integration called the ***EIQ Server***. Their approach is depicted in Figure 4. The EIQ Server approach assumes control by:

- Leaving data in the data sources
- Building and maintaining indexes based on the underlying source data schemas or on result-sets from data source systems
- Imposing a rigorous role-based access control system from user-level to data source field or column-level (not just row-level)
- Enabling highly flexible data models to be used for query generation and results data integration
- Processing all queries 100% using the EIQ Server indexes – requires no interim tables and no interaction with the data sources
- Retrieving only final result-set data from data sources

The EIQ Server approach requires neither adaptors nor complex data and schema transforms. Furthermore, the approach does not require that the data be moved from the original data sources and it requires no major ETL processing. It provides consistent and multiple indexes across multiple disparate data sources. It can accommodate virtually

any data source – structured, unstructured, and semi-structured. It is simple to add new data sources. The EIQ Server approach does require an index updating process and storage for the indices[5]. However, it is highly flexible, and it can provide near real-time updates and very high performance.

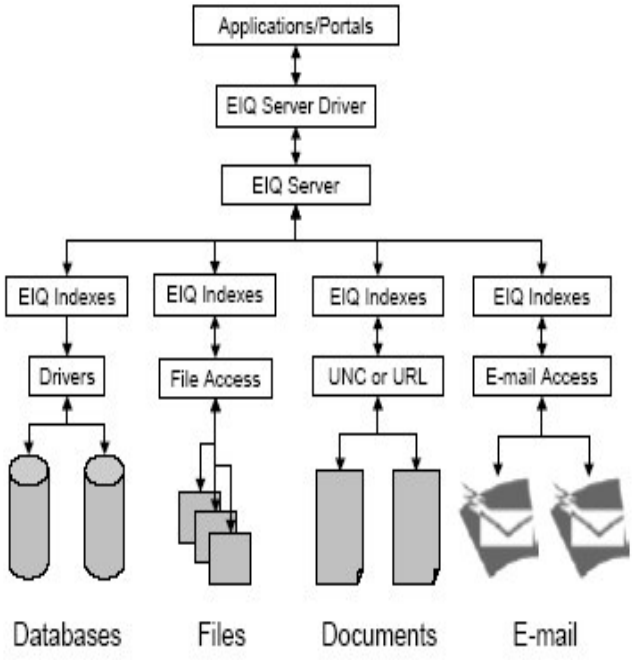


***Figure 4. EIQ Server Approach***

In a nutshell, the EIQ Server combines structured database queries and unstructured text search for almost any data source and for multiple data sources simultaneously. It is deployed as middleware and can be transparent to users and applications on one side, and data sources on the other. The EIQ Server is relatively simple to deploy, as there are no adaptors, no special connectors, and no data schema transforms. The EIQ Server allows multiple index types to be created for data (including cleansed, SOUNDEX, metaphone, stemming, synonyms, and context indexes), and allows total freedom in applying index algorithms and rules for query processing and results retrieval. Finally, since queries are mapped to indexes, the EIQ Server indexes work with multiple metadata dictionaries.

One of the benefits of the EIQ Server in support of the ONA concept is the feature that queries of databases of different formats and structures are possible, regardless of a priori knowledge of the database. The EIQ Server indexes at two basic levels: ***data source schema level*** and at the ***results level***. At the schema level, an Open DataBase

[5] The storage required for the indexes typically ranges from between 20% to 80% of the storage required for the data to be indexed; not the entire database management system) and not necessarily all the data.

Connectivity (ODBC)[6] or other similar connection to the database is required to gather knowledge about the sources prior to indexing the data sources. This requires understanding of the schemas, representation and semantics of each data source.

Results-level indexes, on the other hand, do not require the user to gather knowledge of the data source. Instead, indexes are developed from the results generated from standard queries or reports. For example, a standard Pay Report may produce a table with fields identified as Employee_ID, Total_Hours and Total_Pay. The table contains entries for each employee in the company. The EIQ Server will index the results table. Of course, if the provided results do not provide information about certain fields in the database, the results-level indexing will not know about those data elements. For example, if the data source also contained information about projects and hours charged to each project, the results-level indexing would not support queries about project charges. Nevertheless, results-level indexing can provide very rapid support for some queries against the data source. Results-level indexing usually requires that the database be structured and indexed with some sort of date/time stamp or other indicator that can be used to flag updates or to query against to retrieve final result-set data.

EIQ Server builds and maintains consistent indexes using both levels as appropriate to the database being queried. The end result of a query using EIQ Server is an index that describes the nature of the data available and its location. Only when the information is requested does EIQ Server retrieve the specific final result-set data.

EIQ Server can be described as analogous to a card catalog at a library. Metadata about information contained in the library (books, periodicals, microfilm etc.) is contained in the card catalog which is significantly smaller than the database (library) being indexed. Only when the user decides to retrieve information, does EIQ Server pass the actual data to the user. While the card catalog analogy is useful, it does not adequately portray one of the strengths of EIQ Server – namely, its ability to link information from disparate data sources. To extend the analogy, EIQ Server can be seen as a card catalog of metadata that can also link a book to a recent TV appearance by the author or photographs of the subject.

***Comparison of the Data Integration Approaches***

Table 4 summarizes some of the attributes of EIQ Server compared with the two most common approaches to data integration. The relative rankings are based strictly on the methodologies used in the different approaches – not on actual performance metrics.

Comparison of the advantages and disadvantages of the conventional approaches with the EIQ Server approach suggests ***considering all features, the EIQ Server approach has advantages over the more traditional data warehouse approach with its data transfer issues and the federated database approach with its less flexible adaptors.*** Certainly,

[6] An ODBC inserts a middle layer, called a database driver , between an application and the database management system (DBMS). The purpose of this layer is to translate the application's data queries into commands that the DBMS understands.

the data warehouse and federated database approaches each have their role in some applications, and each is probably optimal under certain conditions. However, the EIQ Server approach seems to provide unique and desirable capabilities under the conditions common to those expected when developing an ONA:

- data cannot be moved
- systems are unable to support queries well
- implementation time (and/or cost) is critical[7]
- the disparate nature of the data precludes a "one size fits all" approach
- the data sources are sensitive to the query load
- the data sources change frequently
- access is required of multiple data sources on multiple platforms in multiple locations

*Table 4. Comparison of Data Integration Approaches*

| Feature | EIQ Server | Federated Database | Data Warehouse |
|---|---|---|---|
| Supports Structured DB Queries | ■ | ■ | ■ |
| Supports Unstructured Text Searches | ◒ | □ | □ |
| Data Stays at Source | ■ | ■ | □ |
| Near Real-Time Index Updates | ◒ | ■ | □ |
| ALL Data Sources – Structured, Unstructured and Semi-Structured | ■ | □ | □ |
| Supports Merging of Results | ■ | ◒ | ■ |
| Use of Original Schemas | ■ | ■ | □ |
| Supports Ad Hoc Queries | ■ | □ | □ |
| Supports Range Queries | ■ | ◒ | ◒ |
| Data Processing Required | ◒ | ■ | □ |
| Transparency of Data Sources to Apps | ■ | ◒ | ■ |
| Additional Storage Required | ◒ | ■ | □ |
| Impact on Data Source | ■ | □ | ■ |
| Non-Intrusive System Changes | ◒ | ■ | ◒ |
| Requirement for Adaptors | ■ | □ | ■ |

Legend

■ ◒ □

Best ... Worst

[7] WhamTech claims that EIQ Server queries are typically 10 to 100 times faster for result-set isolation than other query technologies.

## *ONA Data Integration Pilot Study*

Because of the promising capabilities of the EIQ Server approach to data integration, Project Alpha, in partnership with the ONA team and WhamTech, conducted a pilot study to verify the capabilities and to demonstrate the WhamTech approach.

The ONA team recommended a collection of ONA data sources for the demonstration. Five (5) data sources were used in the pilot study.[8] The data sources were chosen to represent a variety of formats (XML, SQL, Excel), structured and unstructured data, and database update latencies ranging from static to near-real time. All databases were unclassified. The heterogeneous collection of databases used in the pilot study is listed below:

1. ***ONA SQL Server relational database and associated Word and PowerPoint documents***. An unclassified section of the ONA SQL Server database, and PowerPoint and Word documents were provided for the pilot study. The database and documents combine intelligence information from multiple sources and link "nodes" (entities) and effects as the result of applying resources to accomplish diplomatic, information, military or economic (DIME) actions. The database addressed the Far East area in general, and Indonesia, in particular. For the pilot study, the ONA data source was split into two parts: (a) a structured relational database and (b) an unstructured documents database. Both were static databases. The Word documents associated with the ONA database were parsed and indexed as unstructured text. Since most of the information in the PowerPoint files was also available in the Word files in greater detail, the PowerPoint files were not parsed and indexed.

2. ***TRACES.*** The TRACES data source was a single Excel spreadsheet consisting of identity-stripped Patient Medical Records (PMRs). The EIQ Server treated the Excel spreadsheet data as a semi-structured, single-table static database accessible through an ODBC driver for parsing and direct data retrieval. The spreadsheet contained the following data:

    - Patient Identification, including age, service, unit and grade
    - Action in which injury was received
    - Treatment Facility Information
    - Destination Facility Information
    - Injury Description
    - Treatment Description
    - Equipment Used in Transportation
    - Medical History

[8] The original plan called for five ONA data sources. Eleven data sources had to be reviewed, before five data sources were selected, as access and use restrictions imposed by the owners, or irrelevant data caused the elimination of all but five from the pilot study. Among the excluded data sources were GTN, ACTD Rosetta, Census Data, NGA Fortune Cookie and FBIS Web Site.

The patient ID type and patient ID were used as primary and secondary keys to retrieve specific records.

3. ***SEAS Demonstration Data.*** The SEAS data consisted of results from simulations of biological attacks. The simulation generates results in a proprietary, non-standard XML format. SEAS provided two different demonstration result sets for a series of nine time slices in both XML and Excel spreadsheet formats. Since the XML format was non-standard, WhamTech converted the Excel spreadsheets for the two different result sets to standard XML, and then indexed the XML files directly. These two XML data sets were referred to as SEAS1 and SEAS2. The SEAS1 and SEAS2 XML data were mapped to metadata that contains the following information:
    - Country (SEAS1 and SEAS2)
    - Area (SEAS1 and SEAS2)
    - Public Mood (SEAS1)
    - Health Index (SEAS1 and SEAS2)
    - Mitigation Index (SEAS1 and SEAS2)
    - Immune (%) (SEAS2)
    - Susceptible (%) (SEAS2)
    - Neutral Affiliation (% of total population) (SEAS2)
    - Terrorist (% of Muslim population) (SEAS2)

4. ***Web Documents from ONA-Provided News Web Sites.*** A list of web sites of interest to the ONA team (mostly news sites) was provided to WhamTech. WhamTech spidered, parsed and indexed these web sites, seeking specific information on Indonesia – the area of interest to the ONA team. After a number of these web sites were processed and available for queries, it became obvious that the saved links could not be used to retrieve the original web documents, as the links had expired due to the dynamic nature of these web sites. Subscriptions to the web sites would be required to access archived documents. Thus, instead of relying on live links, WhamTech chose to create a repository (or cache) of previously processed web documents. In an actual application, the live link option could be offered in addition to (or instead of) the repository. In total, about 3,600 documents were processed for the pilot project. They included HTML, Word, Adobe Acrobat PDF, and Excel (unstructured).

5. ***RSS News Feeds***. Really Simple Syndication (RSS) is a lightweight XML format designed for sharing news headlines and the content of news-like sites and other web content. Once information about each item is in RSS format, an RSS-aware program can check the feed for changes and react to the changes in an appropriate way. To demonstrate a real-time update to the indexes for the pilot project, an RSS newsreader was used. The RSS feeds are updated constantly 24 hours a day, 7 days a week over the web and the information is made available to anyone "subscribing" to them. The RSS feeds can be filtered to provide only predefined information. For the pilot project, the RSS news feeds were restricted to specific key news web sites as selected by the ONA team and were filtered to

provide only information pertaining to 'Indonesia'. The RSS news feeds generally provide only high-level summary information with a link to a web page containing the details. For the pilot project, WhamTech included the following indexed fields: (1) title, (2) link and (3) date.

***Metadata***. WhamTech was provided access to the DoD XML Registry[9], but it proved to be difficult to use for the pilot project. There were approximately 30 separate metadata repositories covering over 30,000 data elements. However, the data elements available through the registry covered only about 25% of the metadata needed for the pilot project.

***Pilot Project Configuration***. The five databases were configured to work with the EIQ Server as shown in Figure 5. Due to time and resource constraints, only a simple HTML web browser interface was developed for the pilot project. The EIQ Server connection included the five data sources and the indexes that were created by the EIQ Server.

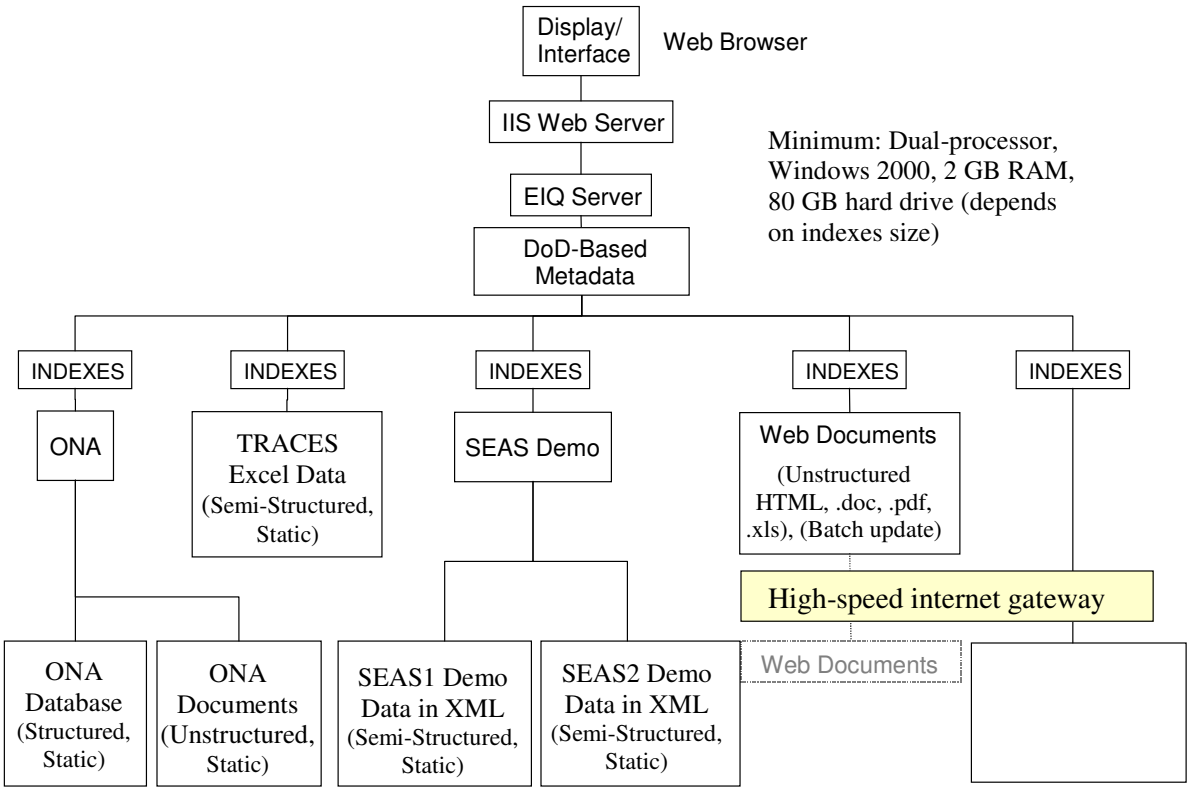


***Figure 5. Configuration for ONA Pilot Study of EIQ Server Approach***

[9]DoD XML Registry – http://diides.ncr.disa.mil/xmlreg/user/index.cfm.

***Demonstration of Queries and Result-Sets.*** WhamTech developed a variety of SQL queries in a drop-down list to demonstrate the functionality of the EIQ Server. In addition to the drop-down list, free format SQL queries requested by members of the ONA team were also demonstrated. The specific data sources provided to WhamTech for use in the pilot study precluded the demonstration of queries that were interesting and useful to the ONA analysts. Nevertheless, WhamTech did demonstrate the ability of the EIQ Server to process a variety of queries (simple and complex) and result-sets from the five data sources. They also demonstrated near real-time index updates on the RSS news feeds.

***Query Response Times***. Although the demonstration was not designed to produce metrics on query response times, the response delays all seemed to be in the sub-second to seconds range. This is consistent with the claims of WhamTech: "*Query times will probably always be in the sub-second to few seconds range. This is true for structured data and unstructured text queries. We have a billion-record database online (www.billionrecords.com) that returns a four-term query in sub-seconds. Range queries and more complex queries can take longer, depending on the size of the interim and final result-sets.*"[10]

***Indexing.*** Typically, certain tasks must be performed by user organizations for data integration independent of the particular approach. For example, the following tasks are all necessary for (a) developing a data warehouse, (b) creating a database federation, (c) developing an XML-based information exchange among systems with divergent schemas, or setting up an EIQ Server with data source schema level indexes:[11]

- *Gather knowledge about sources – prior to integrating multiple data sources, one must understand the schemas, representation, and semantics of each data source*
- *Gather knowledge about desired consumer views – similar to the first task, but this time for the interfaces to be used by consumer users and systems*
- *Identify semantic correspondences among sources and from sources to the consumer views – determine entities and attributes in the different systems that refer to the same (or similar) real world concepts – e.g., that EmpSeniority in a source system can be used for WorkerYearsOfService in the consumer view*
- *Create needed attribute transformations – produce executable functions that transform attributres in the sources to properly feed the consumer views. E.g. specify that one must multiple HourlyWage by HoursWorked to produce Salary.*
- *Specify data combination rules – when multiple source rows each contribute values to a single target row, how should the combination work?*
    - *Join or unior? If a join, on what fields? Inner or outer join?*
    - *What result to produce when sources differ on the same fact?*
- *Create logical mappings from sources to consumer*

[10] Based on email correspondence from Mr. Gavin Robertson, Chief Technology Officer and Senior Vice President of WhamTech, Inc.

[11] Seligman, Len, Arnon Rosenthal, Paul Lehner, Angela Smith. "Data Integration: Where Does the Time Go?", Data Engineering Bulletin, IEEE Technical Committee on Data Engineering, September 2002.

- *Data cleaning – discovering and correcting incorrect data values*
- *Create and optimize an executable connection for the specific run-time environment*

***The EIQ Server data source schema indexing*** requires the tasks listed above. Higher-level ODBC drivers do provide functions that allow data schema discovery, and even associated metadata, attributes, and relationships. Furthermore, there are data profiling tools that can assist with the indexing task. These tools may allow database administrators to quickly learn and map data source structures. However, data source schema indexing still requires human understanding and intervention. Therefore, data source schema level indexing requires, at best, hours to days.

Once the above human tasks have been accomplished, the actual machine indexing proceeds very rapidly. The indexing times depend on the system hardware, the number and length of fields, and the cardinality of the database. However, WhamTech claims the following typical ranges.

"*For web pages, we can parse and index at 10-15 pages per second. For structured databases, the following applies:*

- *100,000 records: 30-60 seconds (3333 to 1667 records per second)*
- *1,000,000 records: 60-120 seconds (16,667 to 8,333 records per second)*
- *100,000,000 records: 20-30 minutes (83,333 to 55,556 records per second)*

*More complex indexes and/or multiple indexes will take longer. If near real-time updates are implemented, index updates should happen almost immediately (sub-seconds), unnoticed in the background."*[12]

As mentioned earlier, the EIQ Server approach indexes at two basic levels:

1. Data source schema level
2. Results level

***Results-level indexing*** avoids the need to gather knowledge about the data source and, therefore, eliminates the need for some of the above tasks. Results from typical "show me the latest data" type queries yield results-level data that can be indexed and queried by the EIQ Server. Examples of situations for which results-level indexing is appropriate (or maybe the only option) are:

- An older legacy system that has very simple index and query processing capabilities
- A very complex or unknown data structure that does not lend itself to being easily indexed
- An uncooperative and possibly hostile data source owner
- A remote data source that is accessible only through a query interface
- A proprietary or secret data source schema

[12] Based on email responses from Gavin Robertson, CTO and Senior VP of WhamTech, Inc.

***Advantages and disadvantages of the two types of indexing***. The data source schema level indexing supports a full range of query possibilities and more immediate updates to indexes. However, it takes longer and is more complex to setup as it requires understanding of the data source schemas and the installation of an index update mechanism. On the other hand, the results-level indexing is faster and simpler to implement since it requires no understanding of the data source schema. However, it is restricted in terms of the queries that can be processed. Nevertheless, results-level indexing is the quickest route to getting indexes for at least some of the data in a data source and supporting queries.

***Preparation for web sites and documents***. If entity extraction options are already in place, little preparation is needed for web sites and documents. All that is needed is for specific web sites to be included in the web spider cue, or to allow the spider to build its own cue from gathered hyperlinks.

***Security Features of the EIQ Server***

The demonstration was conducted using only unclassified and open source databases. The EIQ Server software ran on a government provided computer compliant with DISA Field Security Office (FSO) Security Technical Implementation Guides (STIG). The EIQ Server encountered no difficulties in operating on a STIG compliant machine, demonstrating its compatability with government hardware and software standards.

The development of an ONA will almost certainly require access to classified databases. Although not tested in our demonstration, WhamTech claims that the EIQ Server can accommodate the following four major aspects of security and privacy systems:

1. ***Authentication*** – verification that the user is who he claims to be. EIQ Servers currently communicate with other EIQ Servers through sockets, which allow for currently widely accepted Secure Socket Layer (SSL) secure communications, plus, identity authentication systems such as enterprise Public Key Infrastructure (PKI). Future options for identity authentication, such as SOAP/XML standards, can be easily incorporated.
2. ***Security/Privacy*** – ensures that data are not misused or disclosed to unauthorized people, and ensures that personal identities are protected as far as possible. The EIQ Server offers a platform that can accommodate most, if not all, of the aspects of security/privacy through a comprehensive relational database schema that is flexible enough to accommodate future additions, such as biometric data.
3. ***Integrity*** – protects against data being improperly modified or duplicated. A secure system is needed to protect the integrity of system traffic and files located on the system. The EIQ Server can offer additional integrity protection through a very fast 64-bit CRC algorithm that can sample queries, results and EIQ Indexes, to assure integrity.
4. ***Accountability*** – enables an irrefutable means of tracking operations. The EIQ Server provides and actively monitors and analyzes audit logs to assure accountability.

In addition, WhamTech provides security and privacy access profiles that have been built to accommodate multi-level security. Since WhamTech has not yet pursued accreditation for the EIQ Server, testing of their security model was not part of the pilot study.

## *Conclusions*

Developing an ONA requires vast amounts of data that reside in many distributed heterogeneous data sources, many of which have rarely been accessed by military analysts. It is a serious challenge to integrate the data from the disparate data sources to produce the coherent picture of the adversary that the ONA analysts seek. The data integration problem is critical for many military applications, and solution of the problem has been recognized as a high priority within the DoD.

WhamTech's EIQ Server demonstrated a unique approach to data integration that seems to be well suited to the ONA process and, more broadly, to support processes like EBO, Joint Command and Control, development of a CIE, Terrorism Information Awareness, Joint Deployment and Sustainability or Horizontal Fusion that require the integration of data from multiple distributed heterogeneous data sources. The same situation exists in many situations outside the DoD; for example, in the Department of Homeland Security, intelligence and law enforcement communities, and they are considering the EIQ Server for projects.

- On balance, the EIQ Server seems to offer advantages over the data warehousing and federated databases approaches to data integration under the conditions common to those expected when developing an ONA:
    - data cannot be moved
    - implementation time is critical
    - the disparate nature of the data precludes a "one size fits all" approach
    - the data sources are sensitive to the query load
    - need to combine structured, unstructured, semi-structured (including XML) data in a query
    - need to quickly add new data sources and applications as conditions change
    - need to query data across multiple disparate data sources on multiple platforms in multiple locations – simultaneously

- Despite some artificial constraints, the EIQ Server was able to integrate disparate data sources in real-time. The ability to do this, without time consuming database federation before hand, represents an opportunity for ONA analysts to focus on analysis rather than data and information gathering.

- The results-level indexing capability of the EIQ Server is particularly attractive because it eliminates the need to gather significant knowledge about the data sources (schemas, representation, and semantics). It is therefore faster and

simpler to implement. Results-level indexing seems to provide the quickest route to getting indexes and supporting queries. In some cases (for example, when a data source has an uncooperative and possibly hostile owner, or a proprietary data source schema, or when a remote source is accessible only through a query interface), it may provide the only feasible way to support queries.

- ***Advantages of the EIQ Server Approach to the ONA Analyst.*** The real benefits of the EIQ Server approach to the ONA analyst go beyond time savings. They come from what the EIQ Server allows the analyst to accomplish in addition to the normal processes. These include ***automatic***:

    - Alerts on specific topics, people, places, etc. sent to subscriber individuals or groups
    - Linking of data and information to reveal or discover obvious and non-obvious relationships
    - Creation of searchable archives for future work (web site content changes over time)
    - Presentation and formatting of disparate data and information
    - Connection of ONA systems to systems for EBO, CIE and other defense and interagency systems
    - Support for spanning multiple communities of interest (COIs)
    - Characterization of information

- The DoD is developing many different COI metadata (XML) repositories and is counting on them to improve information interoperability. However, the pilot project exposed several difficulties with using metadata repositories that could degrade the efficiencies promised by data integration, regardless of the approach selected to perform data integration.

    - There are no metadata definitions for most of the data required by ONA analysts
    - There is no ONA-specific metadata repository
    - It is unlikely that any existing or planned COI metadata repository will meet the needs of the ONA analysts. In the pilot study, the data elements available through the registry (across all of the COIs) covered only about 25% of the metadata needed.

## *Recommendations*

- The novel nature of the EIQ Server warrants further investigation and integration into the ONA process or similar concepts requiring the integration of large amounts of disparate data from multiple sources. We recommend that the EIQ Server be used in future experiments involving the ONA with full access to the data sources used to actually produce an ONA.

- We also recommend the EIQ Server approach be considered for areas other than

ONA that require the integration of pre-existing information to improve the efficiency and visibility of current processes. The joint logistics community faces a significant data integration challenge to provide total asset visibility and to support the concepts of Joint Force Deployment and Sustainability and Sense and Respond. We recommend that JFCOM's Joint Logistics Transformation Center and the Joint Deployment Process Owner explore the EIQ Server for the data integration problems that they face.

- We recommend that Assistant Secretary of Defense for Network and Information Integration include the EIQ Server as part of the Horizontal Fusion Portfolio Initiative and be considered for inclusion in a future Quantum Leap proof-of-concept experiment.

- We recommend that DoD consider the addition of an ONA COI metadata repository.

- We recommend that WhamTech seek accreditation of their approach so that the EIQ Server can be used with classified databases.

- Due to time and resource constraints, WhamTech developed only a simple HTML web browser interface for the pilot project. As a consequence, query results were displayed as tables. As a companion pilot study, Project Alpha demonstrated the use of visualization systems to enhance understanding of complex data. We recommend that visualization applications like Starlight,[13] ThinkMap,[14] and Analyst's Notebook[15] be combined with data integration tools like the EIQ Server to provide ONA analysts with a richer understanding of the data available.

## *References*

Stenbit, John. Memorandum of May 9, 2003 on DoD Net-Centric Data Strategy.

Department of Defense Net-Centric Data Strategy, Department of Defense Chief Information Officer, May 9, 2003.

WhamTech, Inc. "Real-time VLDB Database and Search Technologies: WhamTech External Index and Query (EIQ) Server," Version 2.93, October 2003, http://www.whamtech.com/documents/

[13] Starlight™ is a software visualization tool developed at Pacific Northwest National Laboratory to support analysis of complex data and to look for hidden trends and relationships.

[14] ThinkMap™ is a software visualization tool developed by ThinkMap, Inc. to present multiple views of the same data simultaneously and to incorporate multiple data sources into one visual interface.

[15] Analyst's Notebook is a software analysis tool developed by i2 Ltd. It supports timeline analysis, exploration of data relationships and links, and visualization of complex data.

Seligman, Len and Arnon Rosenthal. "A Framework for Information Interoperability," The EDGE: MITRE's Advanced Technology Newsletter, Summer 2004, Volume 8, Number 1, pp 3-4. www.mitre.org/edge

Perez-Nunez, Victor, Robert Jurgens, Larry Hughes, and Ali Obaidi. "Using Data Warehousing to Integrate Multiple Sources of Data," The EDGE: MITRE's Advanced Technology Newsletter, Summer 2004, Volume 8, Number 1, pp. 12, 13, 15. www.mitre.org/edge

Renner, Scott, Dan Hebert, Steve Rainier and John Wilson. "COI Handbook: Practical Guidance for Communities of Interest (COIs) Implementing the DoD Net-Centric Data Strategy," MITRE Technical Report MTR 04B, December 2004.

Miller, Robert W., Mary Ann Malloy and Ed Masek. "Formatted Messaging Modernization Exploits XML Technologies," The EDGE: MITRE's Advanced Technology Newsletter, Summer 2004, Volume 8, Number 1, pp. 18-19.

"Near Real-time Data and Information Integration and Sharing for Project Alpha Rapid Assessment Process (RAP) on Data Integration for the Operational Net Assessment Process," WhamTech Technical Report, June 2004.

Claiborne, Cortney, Darren Cunningham, Davythe Dicochea, Erin O'Malley, Philip On. "Data Integration: The Key to Effective Decisions," MAS Strategies White Paper.

Seligman, Len, Arnon Rosenthal, Paul Lehner, Angela Smith. "Data Integration: Where Does the Time Go?", Data Engineering Bulletin, IEEE Technical Committee on Data Engineering, September 2002.

## *Glossary*

| | |
|---|---|
| CIE | Collaborative Information Environment |
| CIO | Chief Information Officer |
| COI | Community of Interest |
| CRC | Cyclic Redundancy Check |
| DARPA | Defense Advanced Research Projects Agency |
| DIME | Diplomatic, Information, Military and Economic |
| DISA | Defense Information Systems Agency |
| DoD | Department of Defense |
| EBO | Effects Based Operations |
| EIQ | External Index and Query |
| ETL | Extract, Transform and Load |
| FOUO | For Official Use Only |
| FSO | Field Security Office |
| HTML | HyperText Markup Language |
| ID | Identification |
| JFCOM | United States Joint Forces Command |
| NII | Networks and Information Integration |
| ODBC | Open DataBase Connectivity |
| ONA | Operational Net Assessment |
| OWL | Web Ontology Language |
| PKI | Public Key Infrastructure |
| PMESII | Political, Military, Economic, Social, Infrastructure and Information |
| PMR | Patient Medical Records |
| RAP | Rapid Assessment Process |
| RSS | Really Simple Syndication |
| SJFHQ | Standing Joint Force Headquarters |
| SOAP | Simple Object Access Protocol |
| SQL | Structured Query Language |
| SSL | Secure Socket Layer |
| STIG | Security Technical Implementation Guide |
| UNC | Universal Naming Convention |
| URL | Uniform Resource Locator |
| XML | eXtensible Markup Language |