Applying a Unique Approach in a USJFCOM Joint Experimentation (J9) Rapid Assessment Project for Operational Net Assessment (ONA) Data Integration

by
Gavin Robertson, CTO & Sr. VP
WhamTech, Inc.
4450 Sojourn Dr., Suite 200
Addison, Texas 75001
www.whamtech.com

March 2005

WORK PERFORMED UNDER CONTRACT TO THE US GOVERNMENT

TABLE OF CONTENTS

ABSTRACT

WhamTech's unique virtual data integration product called EIQ Server combines the best of federated systems, data warehousing, and enterprise search approaches without many of the associated disadvantages. WhamTech was sourced by USJFCOM J9 Project Alpha to address key questions at several levels, including technical and cultural issues associated with data integration, and ONA processes and data sources: (a) difficulties to overcome, (b) data integration issues, and (c) data structures involved. The project lasted over three months in 2004. WhamTech and Project Alpha selected five data sources out of eleven candidates, comprising (i) ONA SQL Server database and documents, (ii) TRACES medical records in Excel, (iii) SEAS PMESII model output in XML, (iv) Web documents from ONA-specified sites, and (v) RSS live news feeds. The DOD XML Metadata Registry was used for metadata; indexes were developed external to data sources; SQL queries based on metadata including JOINs, range queries, and text search, were executed against the external indexes; result-set pointers were isolated; and results retrieved from data sources. The Project Alpha report, in December 2004, concluded that the approach is unique, has real advantages, could benefit ONA, and recommended extending the work within ONA, JFCOM, and the DOD, and adding visualization.

INTRODUCTION

USJFCOM Joint Experimentation (J9) funded WhamTech to execute a pilot project to explore the unique approach that WhamTech offers to integrate and share structured, unstructured and semi-structured data and information from multiple disparate data sources, and to do so with:

- NO data movement to remote storage
- NO federated system adapters
- NO limits on data sources, indexes, queries (to a certain extent) or metadata
- Almost NO load on existing data source systems

The pilot project had an effective start-date of January 30, 2004, and was scheduled to last about three (3) months. Final installation, testing, presentations and demonstrations occurred during May 25 to 27, 2004.

A WhamTech white paper was issued in June 2004 and the USJFCOM J9 Project Alpha final report was issued in December 2004. The WhamTech white paper focuses on the experiment details, and constitutes to a great extent this paper, which includes:

- Data source selection
- Using and developing metadata
- Index build and maintenance
- Summary and conclusions, including recommendations
- Data source details
- Data source and EIQ Server configuration

- Actual demo queries and results

The USJFCOM J9 Project Alpha final report focuses on higher level issues, such as:

- Data and information integration and sharing issues facing ONA, USJFCOM, and the DOD
- The difference between data exchange and data integration
- The need for semantic interoperability
- Available semantic tools
- Available conventional data integration approaches of data warehousing, federated systems, and enterprise search
- EIQ Server's unique approach and the associated benefits
- Comparison of EIQ Server with conventional approaches
- An independent discussion of the project, including data sources, configuration, query performance, and indexing
- A more general discussion of EIQ Server's indexing capabilities, including results-level indexing, and security aspects
- Conclusions (selected quotes included at the end of this paper)
- Recommendations (selected quotes included at the end of this paper)
- References

Rather than reproducing the USJFCOM J9 Project Alpha final report content, it can be viewed and/or downloaded at www.whamtech.com/documents/USJFCOM%20ONA%20Enabling%20Technology%20Data%20Integration%20RAP%20Final%20Report.pdf.

SCOPE

To evaluate WhamTech's EIQ Server product as a potential solution for data and information integration, sharing, and provision to other applications such as visualization.

OBJECTIVES

To address key questions at several levels, including the overall data and information integration and sharing (hereinafter referred to as "data sharing") technical and cultural issues associated with ONA, the product, WhamTech, and how data sharing fits in with future options for ONA:

- What difficulties are there to overcome?
- What are the data sharing issues?
- What are the data structures involved?

DATA SOURCES

Introduction

Out of eleven (11) possible candidates, five (5) disparate, unclassified data sources were eventually selected:

- DS1 – ONA SQL Server relational database and associated Word documents (structured and unstructured – static)
- DS3 – TRACES, (stripped) patient medical records in an Excel spreadsheet (semi-structured – static)
- DS6 – SEAS Demo biological attack simulation results in two XML files (semi-structured static)
- DS10 – Web documents from ONA-provided news Web sites (unstructured – batch/incremental update)
- DS11 – RSS news feeds, including ONA-provided news Web sites (semi-structured – near real-time)

Data sources that were excluded were GTN database, ACTD Rosetta, Census data, NGA Fortune Cookie, FBIS Web site, and others.

The selection process lasted about a month and data sources were excluded due to:

1. Access difficulties, even though unclassified, they reside on limited access systems
2. Owners were reluctant to allow WhamTech to parse and index content; not necessarily a copyright issue, but more of a process issue
3. Data was so disparate that it was difficult to see any commonality with other data sources

The selected data sources are described in detail as follows:

ONA – DS1

An unclassified section of the Operational Net Assessment (ONA) SQL Server database, and PowerPoint and Word documents were provided to WhamTech for a region of Asia, in general, and the country of Indonesia, in particular. The data source consisted of a structured relational database (DS1a) and unstructured documents (DS1b).

The ONA database and documents contain intelligence analyst information read and entered from multiple data sources. The database is highly structured in linking “nodes” (entities), actions, and effects at a high level, but highly unstructured in descriptions of nodes, actions, and effects. Classification levels and reliabilities are applied to all data. Nodes are key entities involved, including organizations, places, and people. Actions consist of conventional actions: diplomatic, information, military, and economic (DIME). Effects consist of conventional effects: political, military, economic, social, infrastructure, and information (PMESII), as well as unconventional effects: coercion, compulsion, deterrence, defeat, and transition.

Using an ODBC driver command, WhamTech obtained schema information from the unclassified section of the ONA database and deduced the basic data model. Some tables were empty and others were not directly connected to the core set of tables. Another link table had parent-child relationships whose context was not totally understood, and was

therefore excluded from the EIQ Server relational index, although it could be added later. See Figure 1 for a diagram showing the EIQ Server relational index for the ONA database.

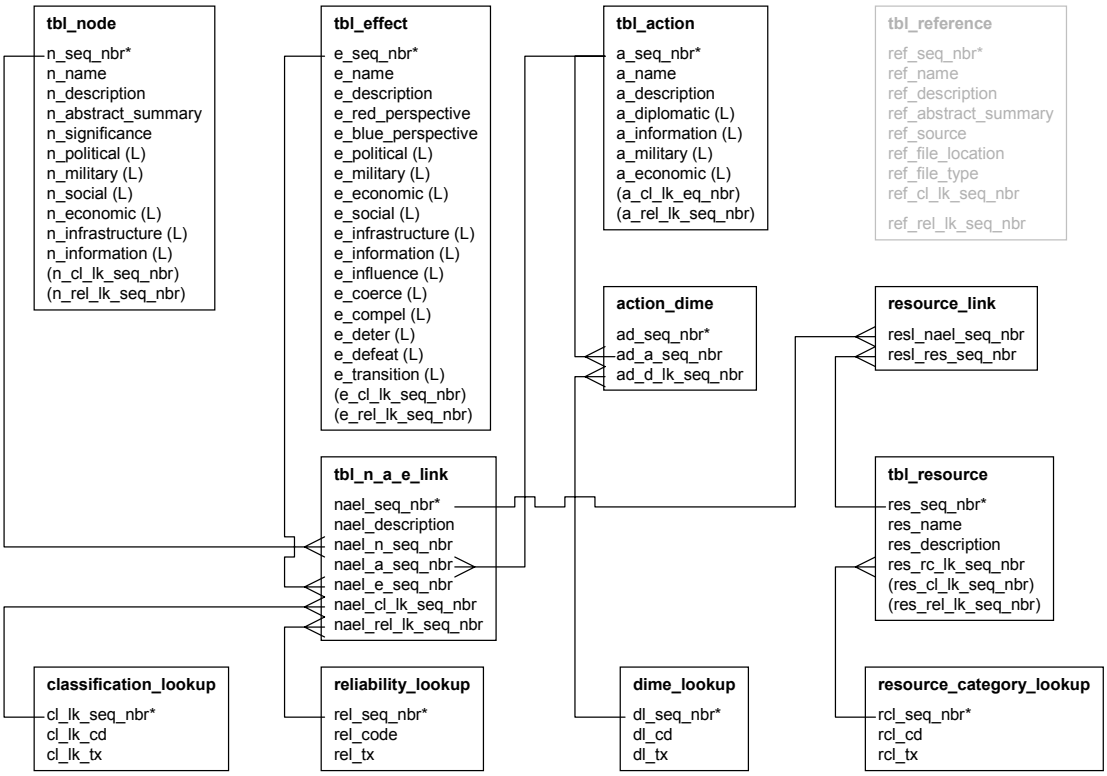


Figure 1: EIQ Server relational index for the ONA database

Appendix 1 contains detailed information on the ONA database and the metadata used for the ONA database.

The unstructured text within the ONA database and Word documents associated with the ONA database were parsed and indexed as unstructured text. The PowerPoint files were not parsed and indexed, as most of the information appeared to be available in the Word files in greater detail. The PowerPoint files could be parsed and indexed if needed.

TRACES – DS3

The TRACES data source is a single Excel spreadsheet consisting of identity-stripped Patient Medical Records (PMRs). Basically, there are three ways to treat data and information in a spreadsheet:

1. Unstructured text: parse and index similar to a document, little or no context to the data and any links would just open the spreadsheet
2. Structured data: treated as a tables or tables in a database, accessible through a driver for parsing and direct data retrieval
3. Hybrid unstructured text/structured data: unstructured text with structured metadata describing the row and column in a spreadsheet where data resides; no direct access to data, as any links would just open the spreadsheet, but would be able to inform you in which row or column the data resides

WhamTech chose to go with option 2, where EIQ Server treats the Excel spreadsheet as a structured, single-table database.

The spreadsheet contains the following data in summary:

- Patient identification, including age, service, unit, grade
- Action in which injury received
- Treatment facility information
- Destination facility information
- Injury description
- Treatment description
- Equipment used in transportation
- Medical history

The Patient ID Type and Patient ID were used as combination primary and secondary keys to retrieve specific records.

As the TRACES data is very patient/treatment specific and the number and types of data sources used in the pilot project are limited, it is therefore difficult to relate TRACES data to other data sources in the pilot project, as there is little overlap or commonality. Much of the indexed unstructured data is full of medical abbreviations and acronyms. Some of the unstructured data would benefit from entity recognition and indexed as structured, e.g., geographic and organization data.

Appendix 2 contains detailed information on the TRACES spreadsheet columns and the metadata used for the TRACES data source.

SEAS Demo – DS6

SEAS, LLC have a product or service that simulates biological and probably other types of attacks. The simulation generates a proprietary, non-standard XML format that is used to display simulation results. Through Project Alpha, SEAS, LLC provided WhamTech with two different demo simulation result-sets for a series of nine time slices in both XML and Excel spreadsheet format. As 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 file directly. The XML data sets were called SEAS1 (DS6a) and SEAS2 (DS6b).

The SEAS1 XML data that was mapped to metadata contains the following data:

- Country
- Area
- Public Mood
- Health Index
- Mitigation Index

The SEAS2 XML data that was mapped to metadata contains the following data:

- Country
- Area
- Immune (%)
- Susceptible (%)
- Neutral Affiliation (% of Total population)
- Terrorist (% of Muslim Population)

The outcome is a unique approach to making XML data available to SQL queries in a federated data integration and sharing system. The XML hierarchical schema consists of Country, Area, Time-slice, and results, and was virtually "flattened" and treated as a single-table for the pilot project; however, it is possible to retain the hierarchical schema and execute standard structured SQL JOINs between other data and XML "branches". The index could also capture the XML hierarchical structure as a network, if there is more than one occurrence of the same data in the hierarchical XML structure.

Of the above XML data, only six fields were indexed in each XML file, as almost all the data is very specific to the SEAS demo and not of general interest. If necessary, all of the fields could be indexed.

As the SEAS simulation results data is very specific to the SEAS demo and the number and types of data sources used in the pilot project are limited, it was difficult to relate SEAS demo data to other data sources in the pilot project, as there is little overlap or commonality with other pilot project data. The data was also very structured.

Appendix 3 contains detailed information on the SEAS XML tags and the metadata used for the SEAS data source.

Web Documents – DS10

WhamTech was provided a list of Web sites of interest by ONA; most of which are news sites. WhamTech spidered (a.k.a. crawled), parsed and indexed these Web sites, seeking specific information on the country of Indonesia. Once a number of these Web sites were processed and available for queries, it became obvious that WhamTech could not use the saved links to retrieve the original Web documents, as the links had expired after a short period (12 to 24 hours) due to the dynamic nature of these Web sites, requiring subscriptions and a different link to access archived Web documents. WhamTech therefore chose to create a repository (or cache) of previously processed Web documents, instead of relying on live links. In the future, the live link option could be offered in addition to the repository, similar to Web search engines.

In total, about 3,600 documents were processed for the pilot project include HTML, Word, Adobe Acrobat PDF, and Excel (unstructured).

No other processes, such as entity recognition, summarization, and other Knowledge Management (KM), were applied to the Web documents, although any of these processes and more, could be included, as needed.

RSS News Feeds – DS11

To demonstrate a real-time update to the indexes for the pilot project, an RSS News Reader was used. RSS feeds are updated constantly 24/7 over the Web and are generally a means of making the latest information available to anyone "subscribing" to them. 100s of news services worldwide and 1000s of Blog Web sites now offers RSS feeds.

RSS feeds can be specifically filtered to only include predefined information. For the pilot project, WhamTech limited the RSS news feeds to specific key news Web sites, as provided by ONA, and included a filter on the country of Indonesia.

RSS news feeds are generally very high-level summary information and include a link to a Web page containing the full story. For the pilot project, WhamTech included the following indexed fields:

- Title
- Link
- Date

The link field was not hyperlinked, as similar to the Web documents; the link expires quickly and requires a subscription in many cases, to retrieve the full article. WhamTech chose not to create a repository (or cache) of RSS-linked documents and not to provide a live link. In the future, both options could be offered in addition to the RSS high-level fields.

METADATA

WhamTech was provided access to the DOD XML Registry: http://diides.ncr.disa.mil/xmlreg/user/index.cfm, which is very comprehensive and from a facility point-of-view, a credit to the visionaries who created it; however, there are several aspects that made it difficult to use for the pilot project:

- There were 30 separate metadata dictionaries, each created for a specific division within the DOD and/or a specific purpose
- The largest metadata dictionary, by far, is TBD – To Be Determined, with over 14,000 data elements
- Of the 30,000+ data elements available in the various metadata dictionary, there is a lot of redundancy

Of all the data elements available to WhamTech through the DOD XML Registry, WhamTech was able to use some from COAL, GMI, INT, PER, and TBD; about 25% of the metadata needed for the pilot project; the rest, about 75%, had to be created.

CONFIGURATION

As data sources were identified, they were configured to work with an EIQ Server, as shown in Figure 2. A visualization application was not added due to time/resource constraints, but could be added, as needed.

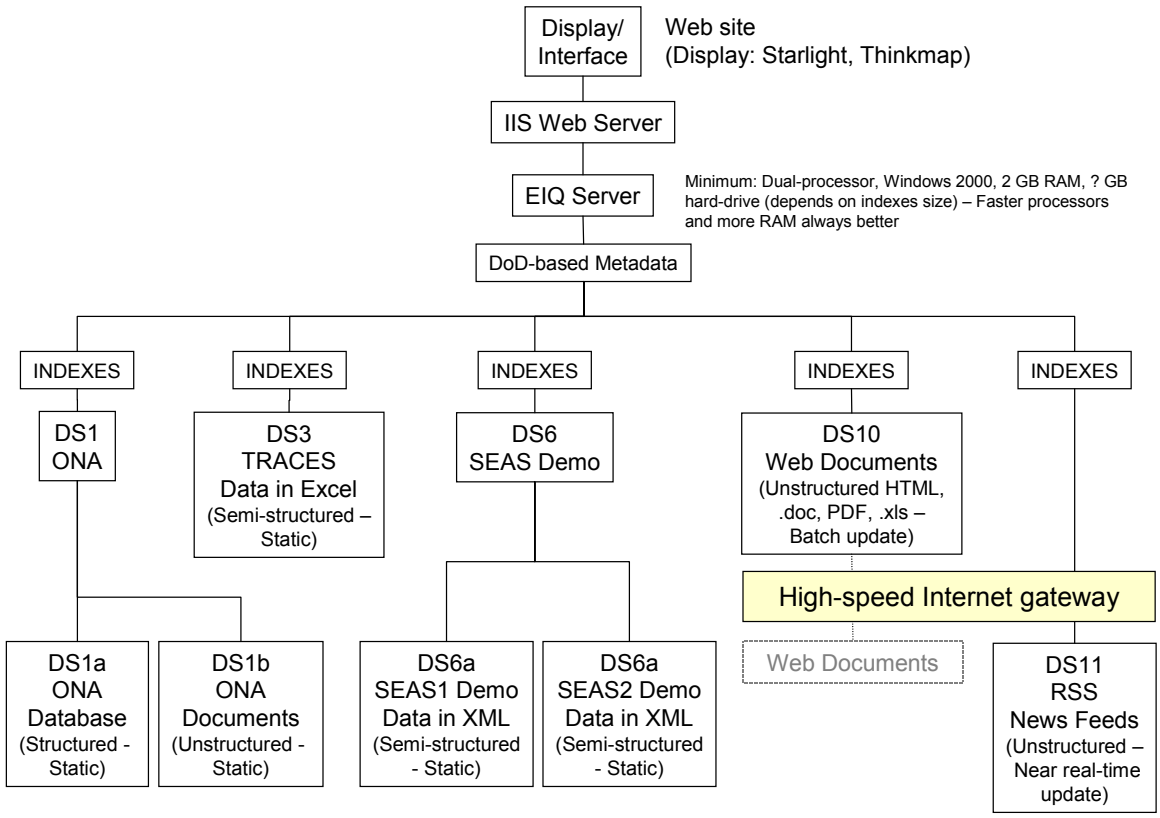


Figure 2: ONA Pilot Project configuration diagram

Configuration Process

EIQ Server has various tools and utilities that readily allow EIQ Server to be configured, most of which can be viewed as screen shots in Appendix 4. The process is generally described as follows:

- Register a data source
- Create a Virtual Data Source (an index and registered data source pair)
- Create a Superschema metadata result-set table containing a list of the data and information of interest
- Map data source fields to Superschema metadata
- Configure EIQ Server Advanced Settings

DEMONSTRATION OF QUERIES AND RESULT-SETS

WhamTech demonstrated an EIQ Server processing queries and result-sets from the five (5) data sources. Also, WhamTech demonstrated near real-time index updates on the RSS news feeds.

A simple HTML Web browser interface was developed for the pilot project, but almost any interface could work with EIQ Server, including existing applications, Communities of Interest (COIs), and complex Java controls in a browser interface. In production deployment, users/applications and data sources would generally not be aware that they are using or being accessed by EIQ Server, as no special connectors, drivers or interfaces are required.

SUMMARY

In summary, the pilot project was developed to answer key questions at several levels on technical and cultural issues associated with data and information integration and sharing, as it relates to ONA. As the project progressed various issues were encountered; some known and others were unknown. In particular, access and cultural issues cannot be underestimated.

The specific key questions defined in the Objectives section are addressed in the Conclusions section. Also, the Recommendations section suggests how the pilot project could be expanded to meet ONA and other DOD requirements.

CONCLUSIONS

- Access to data sources was restricted much of the time. This was mainly due to unclassified data residing on classified or protected networks. It was difficult for WhamTech as an external unclassified contractor to gain access to systems. An operational deployment of an ONA system would probably not encounter these restrictions.
- Cultural barriers to sharing data were less pronounced, but were reflected in a few of the data source owner's responses. This could be understood to a certain extent, as resources are constrained and prioritized elsewhere, and pilot projects take less precedence than operational systems.
- Within the DOD, there are many metadata dictionaries to choose from, with many similar definitions for the same metadata; yet no definitions for most of the data within ONA. The proliferation of metadata dictionaries appears to be the result of no one governing body to manage the process.
- Based on the fact that so many metadata dictionaries exist and will probably continue to exist in the short to medium-term, there is a need to accommodate more than one metadata dictionary for the same data.

- There is a need to accommodate disparate "standard" domain values within ONA and DOD, e.g., E01 instead of E1 for "Enlisted, Grade 1", as each data source could potentially use their own version of so-called "standard data".
- There is a need for more than one metadata mapping WITHIN a metadata dictionary to the same data, allowing a query on a common metadata field, such as Description, to also find the same query term, in the metadata field Summary, for example, without adding a series of OR query constraints.
- As the ONA database system is capable of supporting basic queries, it could probably be indexed and queried using results-level indexing instead of schema-level indexing; this would simplify data access, and work involved in indexing and creating a query execution plan.
- WhamTech took a novel approach with Excel; using ODBC to create an external index of the Excel spreadsheet content; keeping track of the row and column in which data occurs; and using ODBC to retrieve row and column-level data from Excel. In effect, the Excel spreadsheets were treated as single-table databases.
- WhamTech took a novel (and unique?) approach with XML, where XML data was parsed and indexed; treated as a single-table database (although other, more structured options exist); queried using SQL; and using direct file access to retrieve effective row and column-level data from XML. The XML data could be JOINed to other data source data using SQL.

Some of the Conclusions from the USJFCOM J9 Project Alpha final report:

- *"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."*
- *"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 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."*

RECOMMENDATIONS

- There is therefore a need for a consistent and comprehensive global common metadata dictionary for the DOD. It is recommended that such a global common metadata dictionary be developed using the best of the existing 30 metadata dictionaries, and then incrementally developed over time through one governing body to manage the process. Rather than force all organizations to adopt the global common metadata dictionary, it is also recommended that the global common metadata dictionary be develop independently from, and in addition to, existing metadata dictionaries (see a later recommendation for a solution).
- There is a need for an ONA-specific metadata dictionary, or in a possible New World, metadata in the global common metadata dictionary that covers ONA.
- The following recommendation could apply to any data and information integration and sharing system:

  Assuming a data and information sharing system is developed based on a common global metadata dictionary, provision could be made to map data sources AND applications to the common metadata dictionary. Many data and information integration and sharing systems rely on applications being written to a common metadata dictionary, and that may prove difficult in the DOD, as there are already many (at least 30) metadata dictionaries, each associated with a particular organization within the DOD. Even within an organization, there is a dependency on legacy applications, each using their own metadata.

  The recommendation is to consider establishing application and individual organization data source and application mappings and transformations to a common global metadata dictionary. Thus, through combinations of one-to-common and common-to-one mappings, many-to-many mappings are developed, enabling almost any application to work with almost any data source within the DOD.

  To work outside the DOD, e.g., for COALITION forces or integration and sharing with other agencies/organizations, common global dictionaries could be mapped and transformed, extending the many-to-many capabilities.
- In an integrated ONA system:
  - As the ONA data is of very high quality, the ONA system data could be used as the basis to find similar data in other systems in near real-time; perhaps using text-mining software.
  - Other systems could be used to populate and support the ONA system in near real-time; again, perhaps using text-mining software.
  - The use of entity recognition to glean structured data from unstructured text would be an extremely valuable enhancement to the parsing and indexing processes EIQ Server already uses.
  - Other Knowledge Management (KM) tools, e.g., semantic reasoning tools, would provide more intelligent indexing, query processing, categorization, and summarization; allowing analysts to focus on analysis rather than data and information gathering. Plus, this should facilitate a closer to near real-time Assessment -> Planning -> Execution -> Assessment loop.

- Real-time, interactive visualization could add significant value to ONA, as an alternative to static data mart visualization; maybe even a front-end to interactive analysis for the analyst.
- An ONA tool could be developed to tie-in closer to BPM (business process management/modeling) and allow people, processes and technology to be better aligned.
- Multiple ONA systems could be integrated and shared at a higher more user-friendly level with EIQ Server.
- The COI approach could be a great application for EIQ Server and of significant benefit to ONA, overcoming security and cultural barriers, and allowing groups of analysts to collaborate. Similar to ONA systems, COI data and analyses could be integrated and shared across COI repositories.

Some of the Recommendations from the USJFCOM J9 Project Alpha final report:

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

APPENDIX 1: ONA – DS1 DATABASE AND DOCUMENTS DETAILED INFORMATION AND THE METADATA USED

Table and Description (if not obvious)	Column Header	Description (if not obvious from Table and Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
classification_lookup	cl_lk_cd	Code	U, C, R, S	CLASS_LVL	CHAR(1)	Same (GMI)	U, C, S, T (Top Secret)
	cl_lk_tx	Text	UNCLASSIFIED, Confidential, Restricted, Secret	*CLASS_LVL_TX	CHAR(12)		
dime_lookup	dl_cd	Action to be taken/taken as code	D, I, M, E	ACTION	CHAR(1)	Same (INT)	Different – difficult to relate
	dl_tx	Action to be taken/taken as text	Diplomatic, Intelligence, Military, Economic	*ACTION_TX	CHAR(12)		
reliability_lookup	rel_code	Code	N/A, Low, Med, High	EVAL	CHAR(4)	Same (GMI, INT)	Different – difficult to relate
	rel_tx	Text	Not Established, Low, Med, High	*EVAL_TX	CHAR(15)		
resource_category_lookup	rcl_cd	Code	A, G, M, O, Blank	*RESOURCE_CAT_CD	CHAR(1)	C_MA_TY_CD (TBD)	Integers for AIR, SURF, MAR, No "Other" or blank – could be mapped
	rcl_tx	Text	Air, Ground, Maritime, Other, No Selection	*RESOURCE_CAT_TX	CHAR(12)		
tbl_action	a_name	Action name	Various phrases	C_ACT_NM	VARCHAR(150)	Same (TBD)	
	a_description	Action Description	Detailed description	C_ACT_DESCR_TX	VARCHAR(1000)	Same (TBD)	
	a_diplomatic	Action type flag	TRUE, FALSE	*ACTION_D	LOGICAL		

Table and Description (if not obvious)	Column Header	Description (if not obvious from Table and Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
	a_information	Action type flag	TRUE, FALSE	*ACTION_I	LOGICAL		
	a_military	Action type flag	TRUE, FALSE	*ACTION_M	LOGICAL		
	a_economic	Action type flag	TRUE, FALSE	*ACTION_E	LOGICAL		
tbl_effect	e_name	Effect name	Various phrases	*EFFECT_NAME	VARCHAR(150)	Some semblance to ActionEffect (COAL)	
	e_description	Effect description	Detailed description	*EFFECT_DESCR	VARCHAR(1000)	Some semblance to ActionEffect (COAL)	
	e_red_perspective	Effect from RED's perspective	Detailed statement	*EFFECT_RED_PERSP	VARCHAR(1000)		
	e_blue_perspective	Effect from BLUE's perspective	Detailed statement	*EFFECT_BLUE_PERSP	VARCHAR(1000)		
	e_political	Effect type flag	TRUE, FALSE	*EFFECT_POLITICAL	LOGICAL		
	e_military	Effect type flag	TRUE, FALSE	*EFFECT_MILITARY	LOGICAL		
	e_economic	Effect type flag	TRUE, FALSE	*EFFECT_ECONOMIC	LOGICAL		
	e_social	Effect type flag	TRUE, FALSE	*EFFECT_SOCIAL	LOGICAL		
	e_infrastructure	Effect type flag	TRUE, FALSE	*EFFECT_INFRASTRUCTURE	LOGICAL		
	e_information	Effect type flag	TRUE, FALSE	*EFFECT_INFORMATION	LOGICAL		
	e_influence	Effect type flag	TRUE, FALSE	*EFFECT_INFLUENCE	LOGICAL		
	e_coerce	Effect type flag	TRUE, FALSE	*EFFECT_COERCE	LOGICAL		
	e_compel	Effect type flag	TRUE, FALSE	*EFFECT_COMPEL	LOGICAL		
	e_deter	Effect type flag	TRUE, FALSE	*EFFECT_DETER	LOGICAL		

Table and Description (if not obvious)	Column Header	Description (if not obvious from Table and Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
	e_defeat	Effect type flag	TRUE, FALSE	*EFFECT_DEFEAT	LOGICAL		
	e_transition	Effect type flag	TRUE, FALSE	*EFFECT_TRANSITION	LOGICAL		
tbl_n_a_e_link	nael_description	Summary of the "node", action and effect	Detailed description	SUMMARY	VARCHAR(1000)	Same name – different definition (GMI)	
tbl_node	n_name	Node name	Various phrases	C_NODE_NM	VARCHAR(50)	Same (TBD)	
	n_description	Node description	Detailed description	C_NODE_DESCR_TX	VARCHAR(255)	Same (TBD)	
	n_abstract_summary	Node references and content	Detailed statement	*NODE_ABS_SUMM	VARCHAR(255)		
	n_significance	The reason the node is considered significant	Detailed statement	*NODE_SIGNIF	VARCHAR(255)		
	n_political	Attribute flag	TRUE, FALSE	*NODE_POLITICAL	LOGICAL		
	n_military	Attribute flag	TRUE, FALSE	*NODE_MILITARY	LOGICAL		
	n_social	Attribute flag	TRUE, FALSE	*NODE_SOCIAL	LOGICAL		
	n_economic	Attribute flag	TRUE, FALSE	*NODE_ECONOMIC	LOGICAL		
	n_infrastructure	Attribute flag	TRUE, FALSE	*NODE_INFRASTRUCTURE	LOGICAL		
	n_information	Attribute flag	TRUE, FALSE	*NODE_INFORMATION	LOGICAL		
tbl_resource A detailed description of the resource used/to be used – linked to resource category table	res_name			*RES_NAME	VARCHAR(50)		
	res_description			*RES_DESCR	VARCHAR(255)		

APPENDIX 2: TRACES –DS3 SPREADSHEET DETAILED INFORMATION AND THE METADATA USED

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
DB	Database record status	ACTIVE, ARCHIVE	*DB_REC_STATUS	CHAR(6)		
PMR Identifier	Patient Medical Record	Blank	*PMR_ID	INT(8)		
Patient Identifier		Blank	*Patient_ID	INT(8)		
Patient Last Name		Blank	SURNAME	CHAR(30)	Same (INT, PER)	
Patient First Name, MI		Blank	FORENAME+", "+MIDDLE NAME	CHAR(20)+", "+CHAR(20)	Same (INT, PER)	
Patient ID Type		SSN, NATO, Other	PersonIDType	CHAR(36)	Same (PER)	
Patient ID		Blank	PersonID	INT(16)	Same (Many)	
Service Grade Code	Enlisted, Non-enlisted (civilians), Officer, and Warrant Officer	A00, C00, E01 - E09, N00, O01 – O10, W01 – W04, Blank	GRADE	CHAR(3)	CHAR(2) (Many)	E1 – E9, GA – GO, O, O1 – O9, OA, SA – SF, U, W1 – W5, Z
Service Grade Name		Aviation Cadet, Academy Cadet, ROTC Cadet, Dependents, DOD Civilian	*GRADE_NAME	CHAR(54)		

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
		Employees, Contract Surgeons, Enlisted Grade E1 – E9, Officer Grade O1 – O9, "OSI, CID, NIS", Warrant Officer Grade W1 – W4, Blank				
Cite#		Blank				
Current PMR State		Canceled, Completed, Draft, Errored, Executing, Planned, Ready, Suspended, Validated, Blank				
Age		0 - 67	*AGE	INT(2)		
Age Unit		Day, Month, Year, Blank	*AGE_UNIT	CHAR(5)		
Service (Active Duty)		Air Force, Army, Coast Guard, Marine Corps, Navy, Blank	*FORCE_NAME	CHAR(54)	FORCE is used – CHAR(4)	AF, AR, CG, NI, NV
Personnel Status Code		Axy, Cxy, D, Fxy, Hxy, Mxy, Nxy, Oxy, Pxy, Qxy, Xxy				
Personnel Status Name		Air Force Dependent of Retired, etc.				
Unit Name		Mixed	UNIT_NAME	CHAR(54)	Same	

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
					(GMI)	
Movement Classification Code		e.g., A1, with A – F, and 1 – 5, also V9, blank				
Movement Classification Name		Mixed				
Reason Regulated		AF – XP, Blank				
Precedence		Priority, Routine, Urgent, Blank				
Ready Date		M/D/YYYY				
Commercial Transport		N, Y (no blanks)				
# Litter spaces		0 – 6, Blank				
# Ambulatory spaces		0 – 3, Blank				
Origin Facility Code		e.g., A01234, A - V				
Origin Facility Name		Mixed				
Origin Theater		CENTCOM, CONUS, EUCOM, PACOM, SOUTHCOM (no blanks)				
Destination Facility Code		e.g., A01234, A - V				
Destination Facility Name		Mixed				
Destination Theater		CENTCOM, CONUS, EUCOM, PACOM, SOUTHCOM (no blanks)				
Casualty Event		Mixed (limited)				
Injury Type		Battle injury, NDAA, Non-Battle				

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
		Injury, Shrapnel, Sm Px Vac Reac				
Primary MEDSPEC Code		2 – 6 letter acronyms				
Primary MEDSPEC Name		Mixed (limited)				
Primary Diagnosis Code		Numbers and letter coded – treat as char field				
Primary Diagnosis Name		Mixed				
Secondary MEDSPEC Code		2 – 6 letter acronyms				
Secondary MEDSPEC Name		Mixed (limited)				
Secondary Diagnosis Code		Numbers and letter coded – treat as char field				
Secondary Diagnosis Name		Mixed				
Other MEDSPEC Code		2 – 6 letter acronyms				
Other MEDSPEC Name		Mixed (limited)				
Other Diagnosis Code		Numbers and letter coded – treat as char field				
Other Diagnosis Name		Mixed				
Admitted MTF	Medical Treatment Facility	Mixed				
Admission Date		M/D/YYYY H:MM:SS AM or PM				
Last Known ITV Event	ITV?	Mixed (limited)				
Last Known Location		Mixed				

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
Last Known Location Date		M/D/YYYY H:MM:SS AM or PM				
Disposition Date		M/D/YYYY H:MM:SS AM or PM				
Disposition Event		Deceased, Discharged, Return to Duty, Return to Home Station, Scheduled for Onward Movement	DISPOSITION_EVENT	CHAR(50)	Similar (Many)	
Disposition Comments		Freeform text				
# All Attendants		0 - 20				
# MA	Medical Attendants?	0 - 17				
# NMA	Non-Medical Attendants?	0 - 7				
PMR has Itinerary		N, Y				
Proposed Arrival Date		M/D/YYYY H:MM:SS AM or PM				
Mission ID# (of Last Mission)		CHAR(13) max - mixed				
# Missions (PMR Movements)		0 - 6				
# RON Stops		0 – 5				
Cardiac Monitor?		N, Y, Blank				
Incubator?		N, Y, Blank				
Stryker Frame?		N, Y, Blank				
Traction?		N, Y, Blank				

Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
Infusion Pump?		N, Y, Blank				
Ventilator?		N, Y, Blank				
Battery Support Unit?		N, Y, Blank				
Pulse Oximeter?		N, Y, Blank				
Vital Signs Monitor?		N, Y, Blank				
Suction Apparatus?		N, Y, Blank				
Oxygen Analyzer?		N, Y, Blank				
Litter Folding?		N, Y, Blank				
Mattress Litter?		N, Y, Blank				
Straps Webbing?		N, Y, Blank				
Restraint Set?		N, Y, Blank				
Blanket?		N, Y, Blank				
Back Board?		N, Y, Blank				
History		Freeform Text > CHAR(1500)	*MEDICAL_HISTORY	CHAR(2000)		

APPENDIX 3: SEAS –DS6 XML FILES DETAILED INFORMATION AND THE METADATA USED

The SEAS1 XML data set contains the following data:

- Country*
- Area*
- Public Mood*
- Health Index*
- Mitigation Index*
- Released from Shelter (THOU)
- Decontamination
- Exposed (THOU)
- Riot Occurrence
- Demonstration Occurrence
- Government Stability
- Unemployment (%)*
- Cumulative Infections (THOU)
- New Fatalities (THOU)
- Cum. cost of Fatalities ($ MIL)
- Cum. cost of Hospitalization ($ THOU)
- National Monument ICI
- Government office ICI
- Commercial ICI

- Education Institution ICI
- Islamic Institution ICI
- Non-Islamic Institution ICI
- Super Satisfied (%)
- Very Satisfied (%)
- Satisfied (%)
- Neutral (%)
- Unsatisfied (%)
- Very Unsatisfied (%)
- Super Unsatisfied (%)
- Vaccinated (THOU)
- No Resistance (%)
- Low Partial Resistance (%)
- Medium Partial Resistance (%)
- High Partial Resistance (%)
- New Infections (THOU)
- Blue Sympathizer (THOU)
- Neutral Affiliation (THOU)
- Red Sympathizer (THOU)
- Red Passive Supporter (THOU)
- Red Active Supporter (THOU)
- Red Recruit (THOU)
- Terrorist (THOU)

The SEAS2 XML data set contains the following data:

- Country*
- Area*
- National Monument Bombing Occurrence
- Government office Bombing Occurrence
- Commercial Institution Bombing Occurrence
- Education Institution Bombing Occurrence
- Islamic Religious Building Bombing Occurrence
- Non-Islamic Building Bombing Occurrence
- Immune (%)*
- Infected With Symptoms (THOU)
- Infected Without Symptoms (THOU)
- Cumulative Fatalities (THOU)
- Susceptible (%)*
- In Hospital (THOU)
- In Isolation Camp (THOU)
- In Shelter (THOU)
- Blue Sympathizer (% of Total population)
- Neutral Affiliation (% of Total population)*
- Red Sympathizer (% of Total population)
- Red Passive Supporter (% of Total Population)
- Red Active Supporter (% of Total Population)
- Red Recruit (% of Muslim Population)

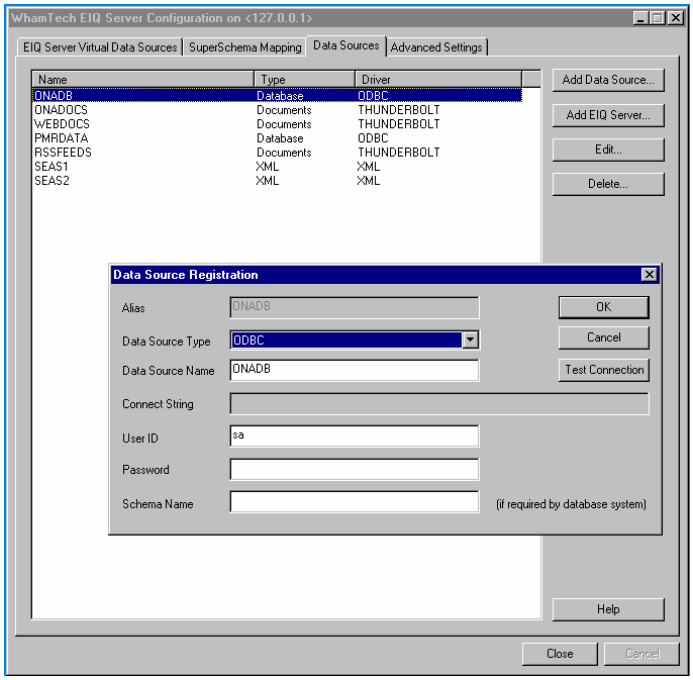
- Terrorist (% of Muslim Population)*
- National Monument Bombing Prevention
- Government office Bombing Prevention
- Commercial Institution Bombing Prevention
- Education Institution Bombing Prevention
- Islamic Religious Building Bombing Prevention
- Non-Islamic Building Bombing Prevention

* = Indexed

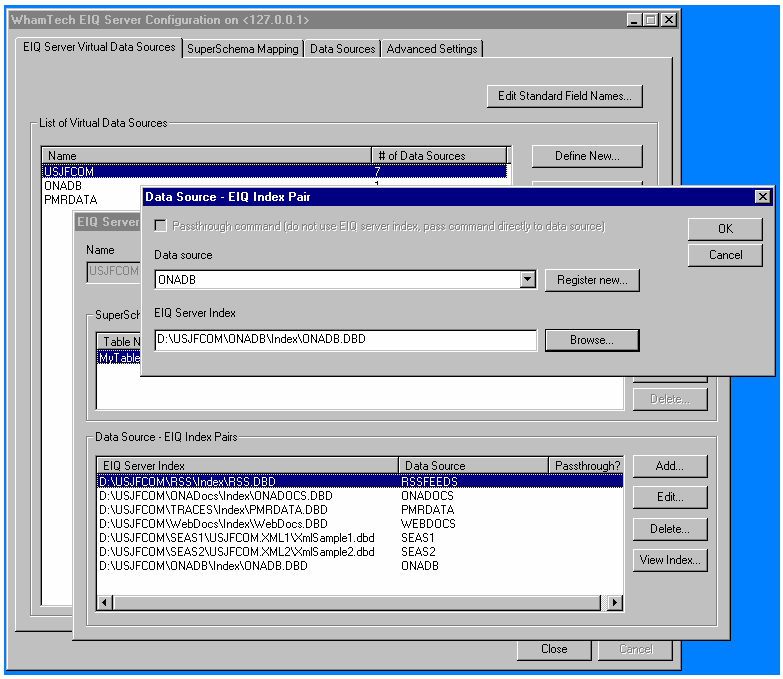
Sheet Name/ Column Header	Description (if not obvious from Column Header)	Domain Value Range	Metadata * = Created	Metadata Attribute	Official Metadata and/or Attribute	Official Metadata Domain Value Range
SEAS 1 & 2/COUNTRY (Sheet)			COUNTRY	CHAR(2)	Same (INT)	
SEAS 1 & 2/AREA (Sheet)			AREA	CHAR(54)	Same (INT)	
SEAS 1/Public Mood	?	2.500 – 4.500	*PUBLIC_MOOD	FLOAT		
SEAS 1/Health Index	?	4.500 – 6.000	*HEALTH_INDEX	FLOAT		
SEAS 1/Mitigation Index	?	29 – 47	*MITIGATION_INDEX	INT(2)		
SEAS 1/Unemployment (%)		10.000 – 14.000	*UNEMPLOYMENT	FLOAT		
SEAS 2/Immune (%)	?	0.000 – 58.000	IMMUNE	FLOAT		
SEAS 2/Susceptible (%)	?	40.000 – 100.000	SUSCEPTIBLE	FLOAT		
SEAS 2/Neutral Affiliation (%)		7.000 – 86.000	NEUTRAL_AFFILIATION	FLOAT		
SEAS 2/Terrorist as a % of Muslims		0.000 – 0.700	TERRORIST_MUSLIM_POP	FLOAT		

Note: As with almost all the data sources, all fields were indexed, but most were not mapped; for a complete list see the body text of this White Paper. Not all the fields in the SEAS XML files were included in the above table; only those that were mapped to metadata.

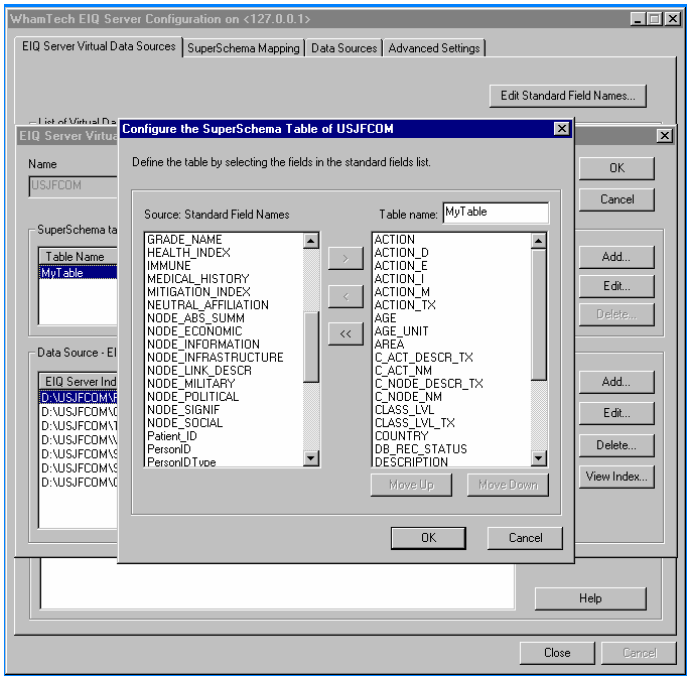
APPENDIX 4: CONFIGURATION SCREEN SHOTS



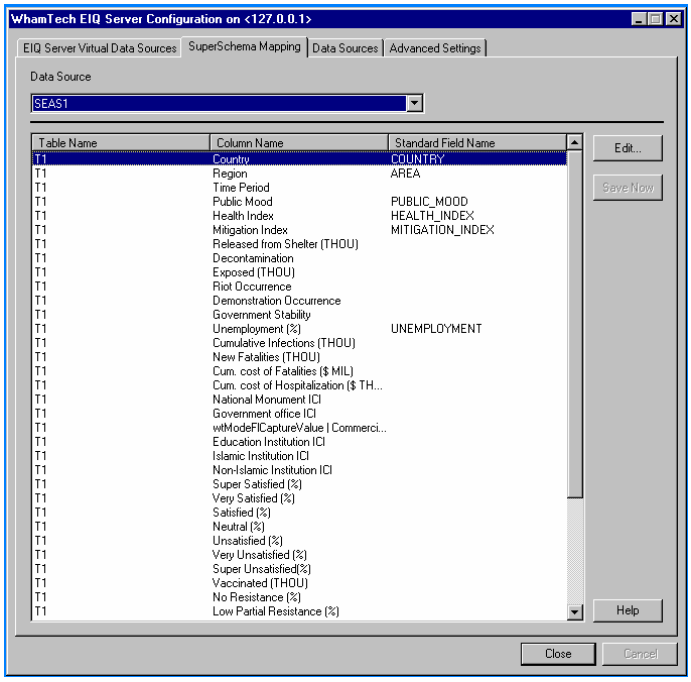
Configuration Screen 1: Registering a data source (the ONA - DS1 SQL Server database)



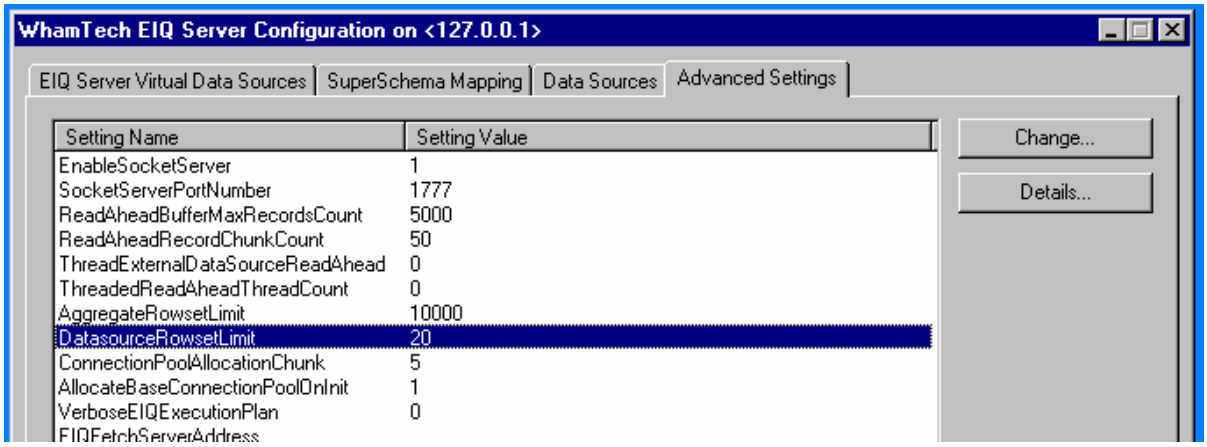
Configuration Screen 2: Creating a Virtual Data Source – an index and registered data source pair (the ONA – DS1 SQL Server database)



Configuration Screen 3: Creating a Superschema metadata result-set table (MyTable, in this case using the ONA – DS1 SQL Server database as a guide)

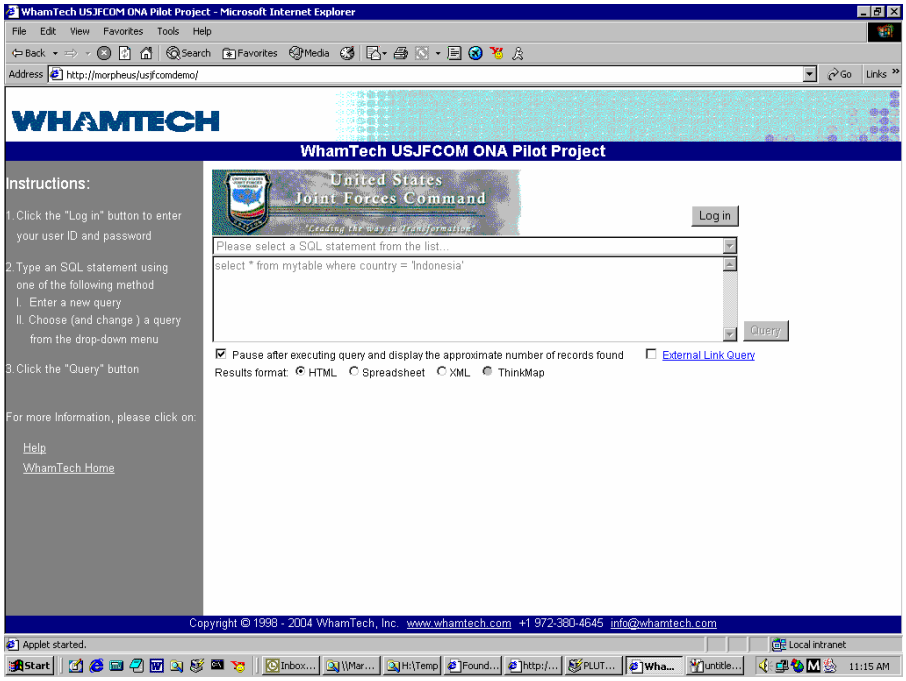


Configuration Screen 4: Mapping data source fields to Superschema metadata (the SEAS1 – DS6a XML file)



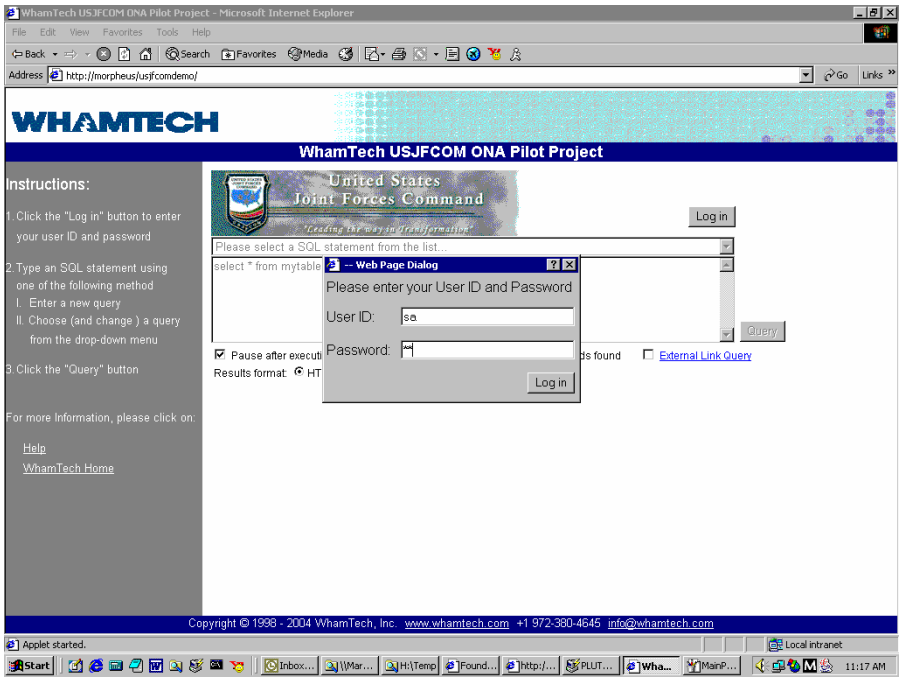
Configuration Screen 5: Advanced settings

APPENDIX 5: DEMONSTRATION QUERIES AND RESULT-SETS SCREEN SHOTS

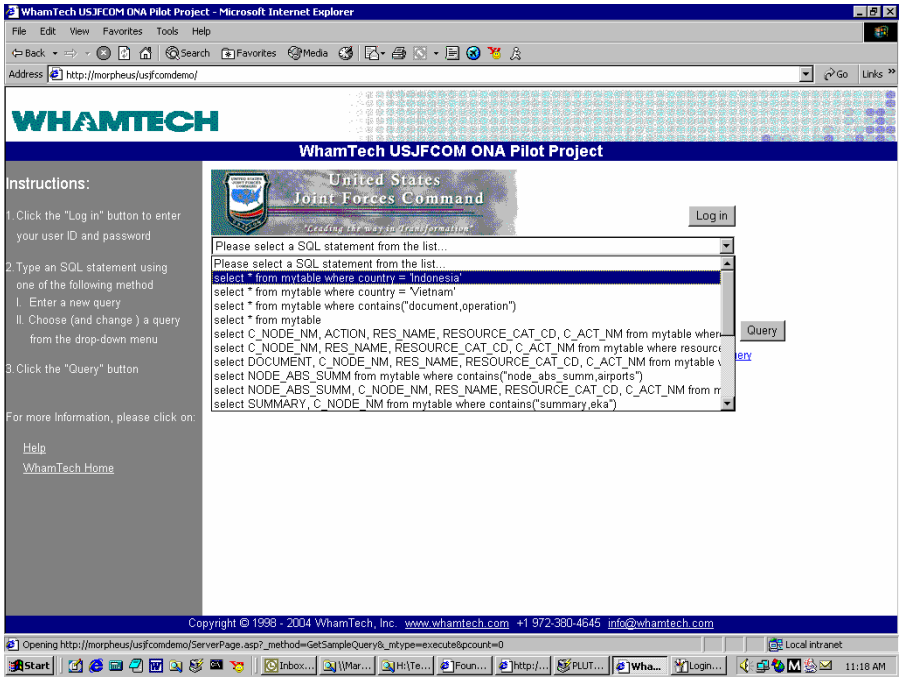


Demo Screen 1: HTML thin-client demo interface - does not need to be so - can be almost any application and/or interface - this interface was built solely to demo the pilot project

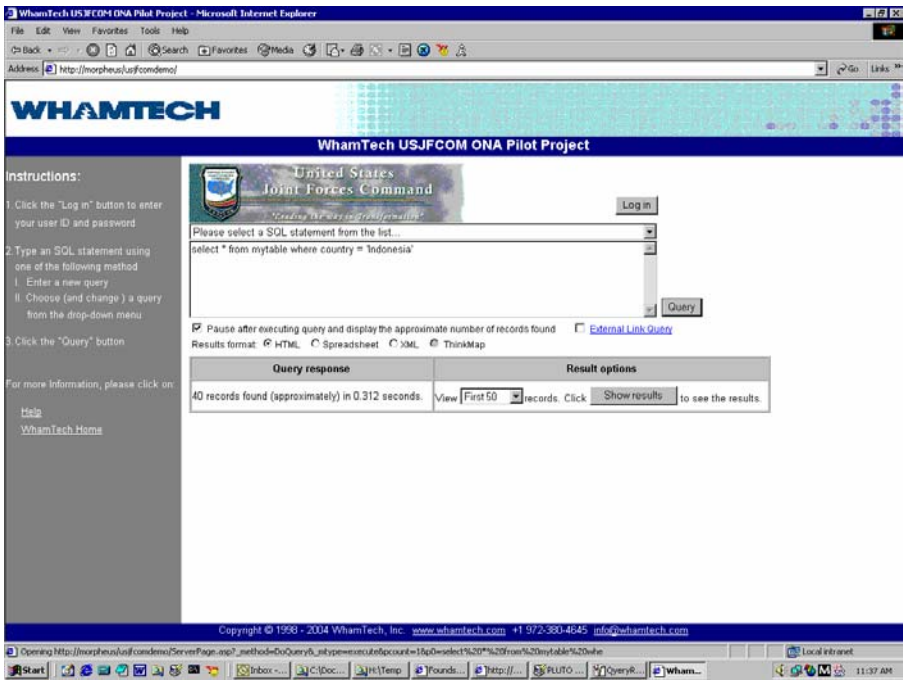
Opening screen automatically connects to an EIQ Server, whose configuration includes the five (5) data sources (seven (7) separate indexes) that are part of the pilot project



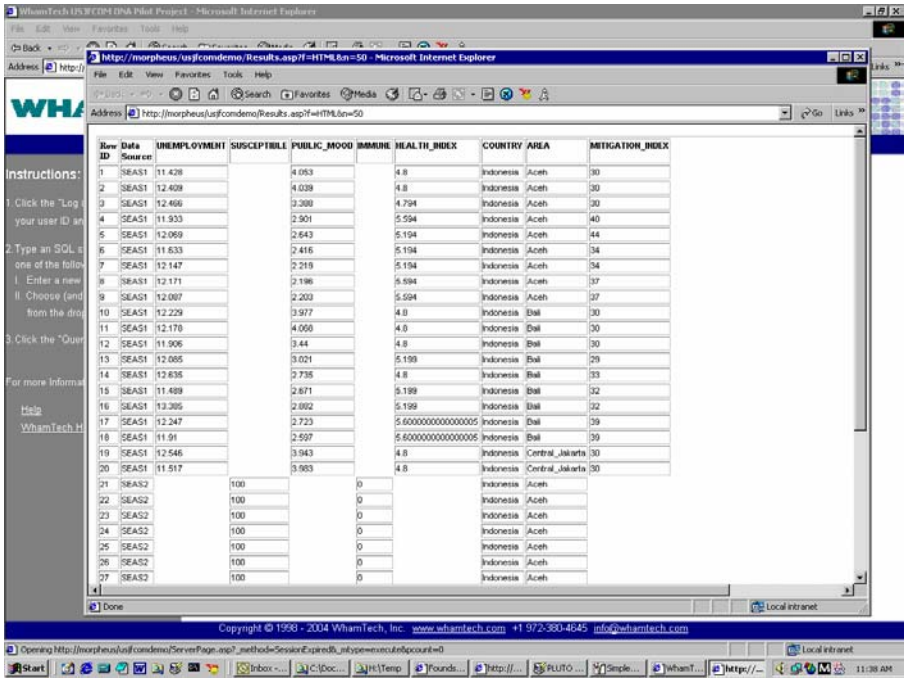
Demo Screen 2: Simple pilot project login. EIQ Server can accommodate much more sophisticated secure ID management



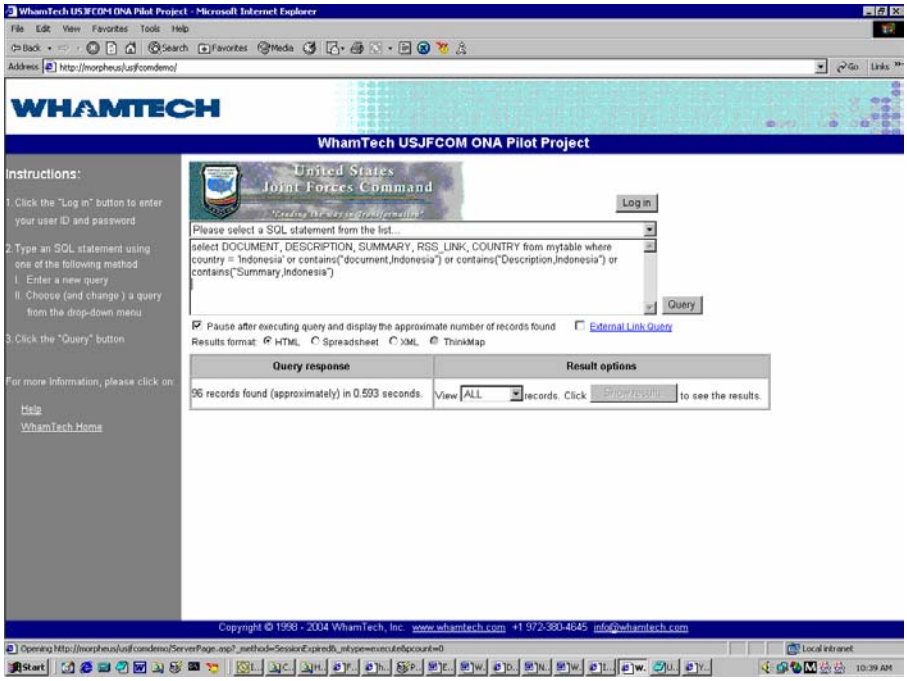
Demo Screen 3: Example list of drop-down SQL queries to save time and provide direction. Queries are totally free format, but should conform to SQL



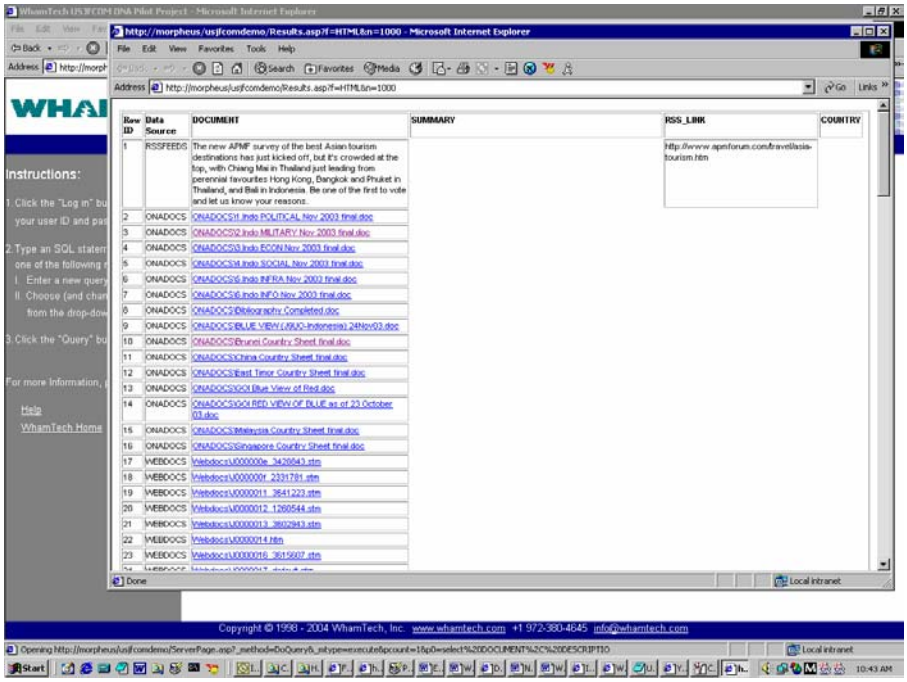
Demo Screen 4: A simple structured data query requesting all available data in the Superschema MyTable where the country is listed as Indonesia - note the option to allow a pause screen to present results metadata BEFORE retrieving results from the data sources



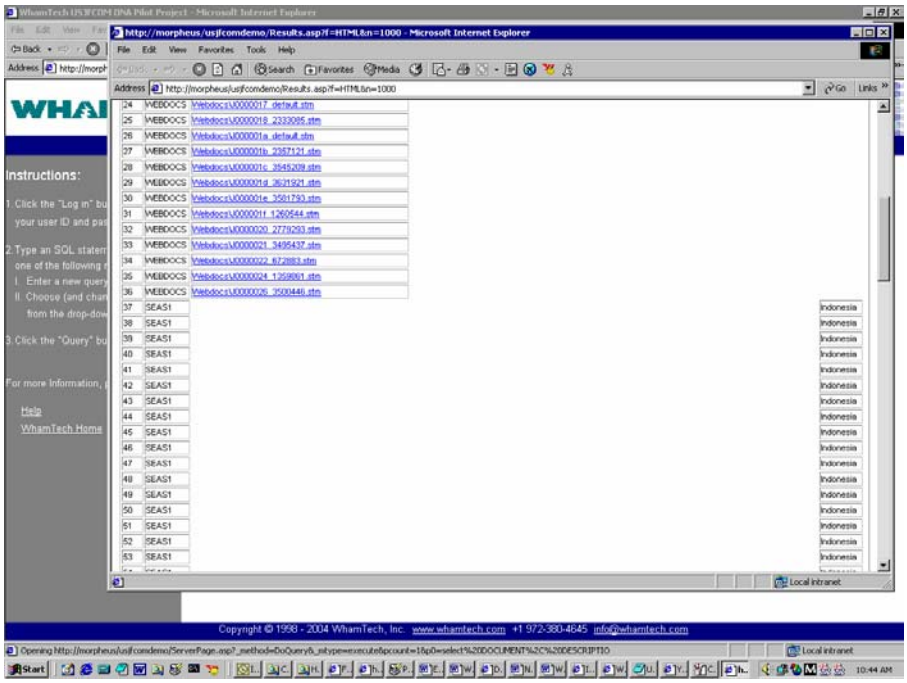
Demo Screen 5: Initial result-set for the query in Demo Screen 4. Note that only the SEAS data source specifically lists a structured field for COUNTRY and therefore the two SEAS XML files were the only ones that produced results. If entity recognition algorithms were used during the unstructured text parsing and indexing, many other results would have been returned



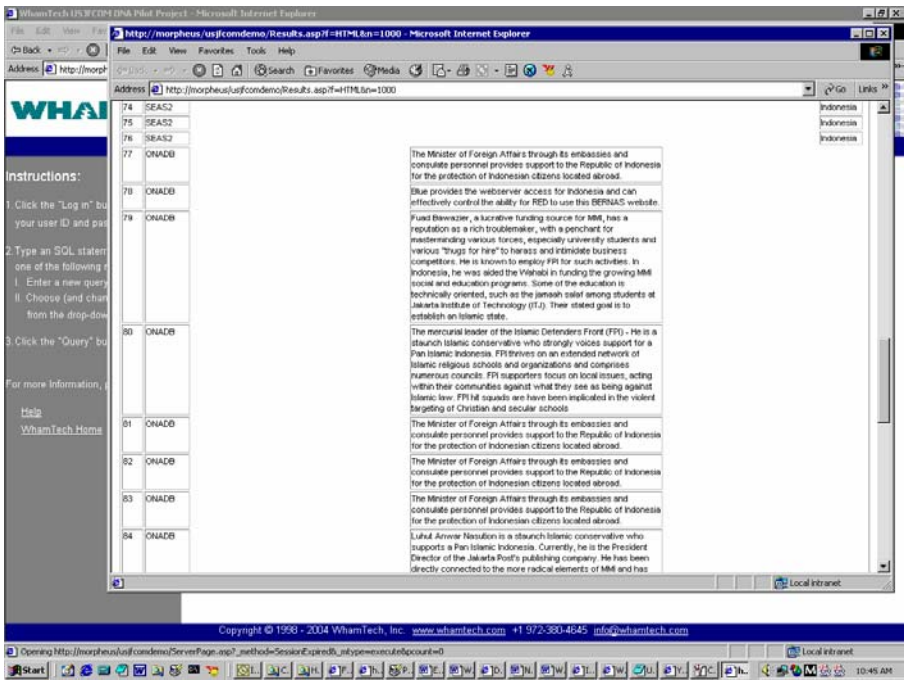
Demo Screen 6: An alternative approach to relying on entity recognition algorithms is to combine a structured data query with an unstructured text query on Indonesia. In this case, to help make results more presentable, the query specified relevant metadata fields from the Superschema MyTable



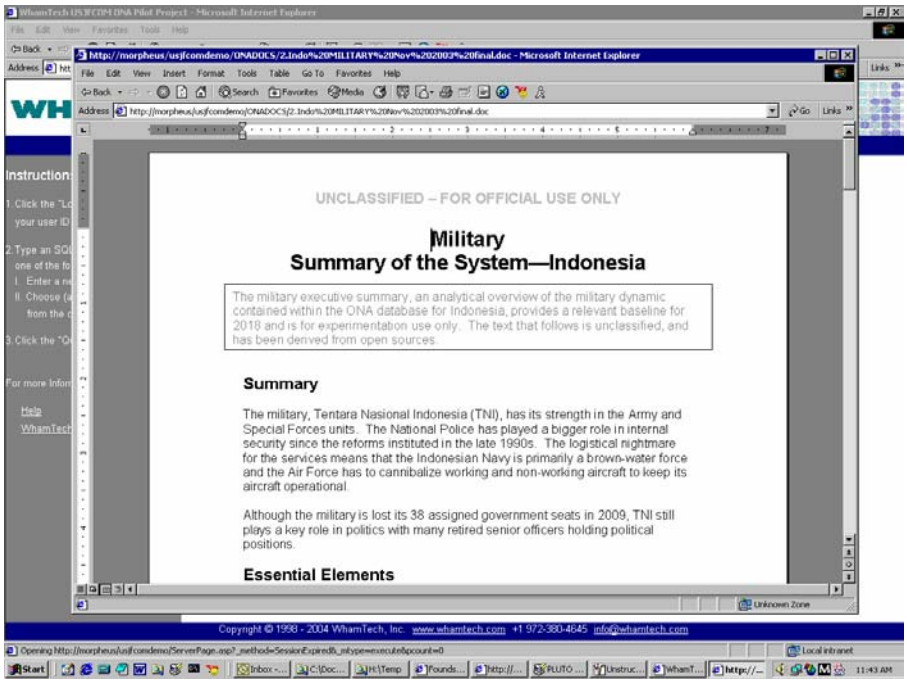
Demo Screen 7: Part 1 of 3 screen shots of the result-set from the combined structured data and text query defined in Demo Screen 6. Almost all data sources contained the query term "Indonesia"



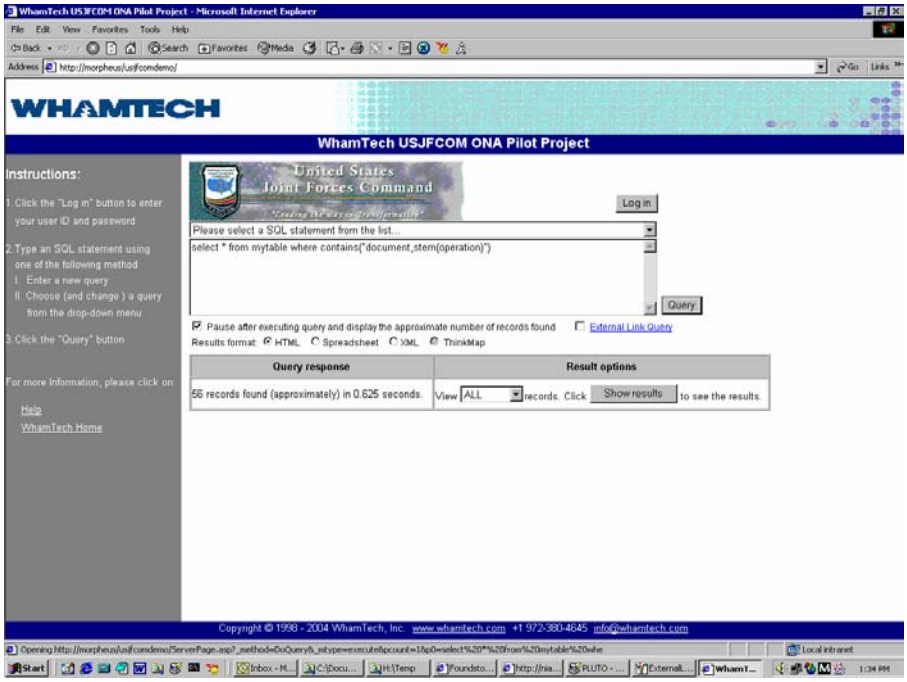
Demo Screen 8: Part 2 of 3 screen shots of the result-set from the combined structured data and text query defined in Demo Screen 6



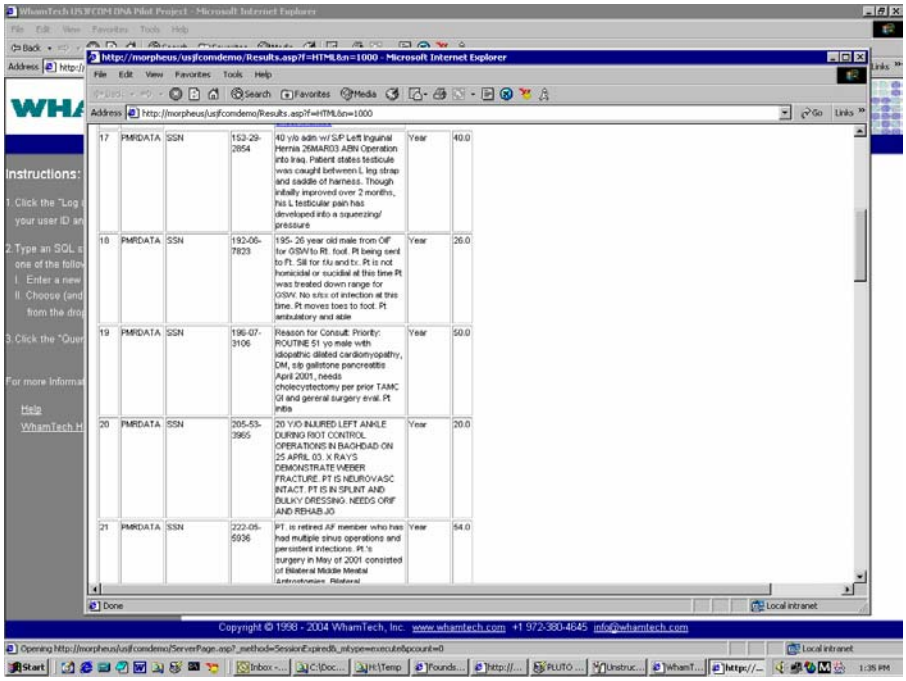
Demo Screen 9: Part 3 of 3 screen shots of the result-set from the combined structured data and text query defined in Demo Screen 6



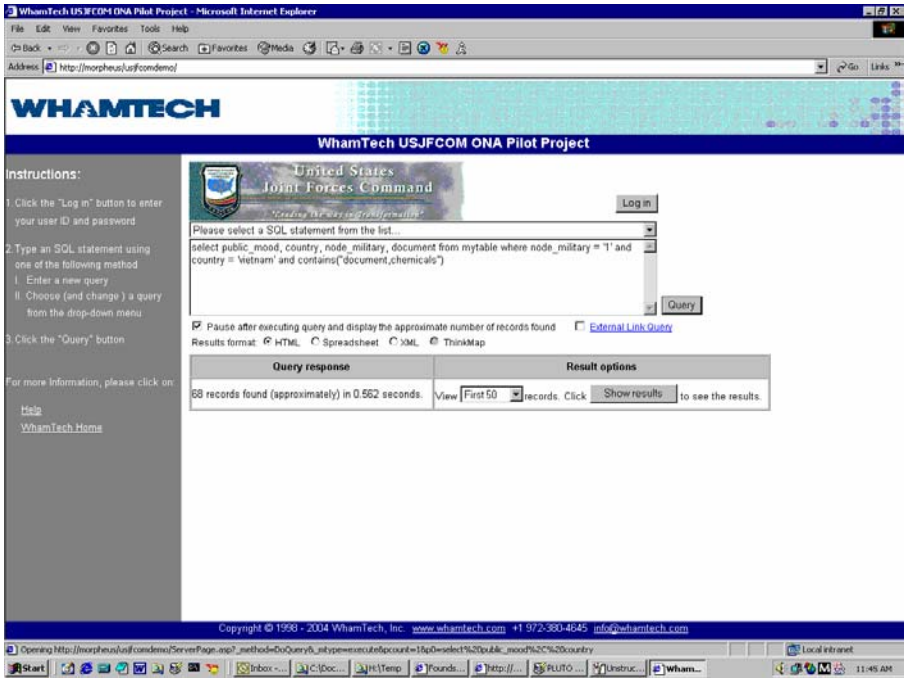
Demo Screen 10: Example of a Word document opened by clicking on a link in a result-set from the combined data and text query defined in Demo Screen 6



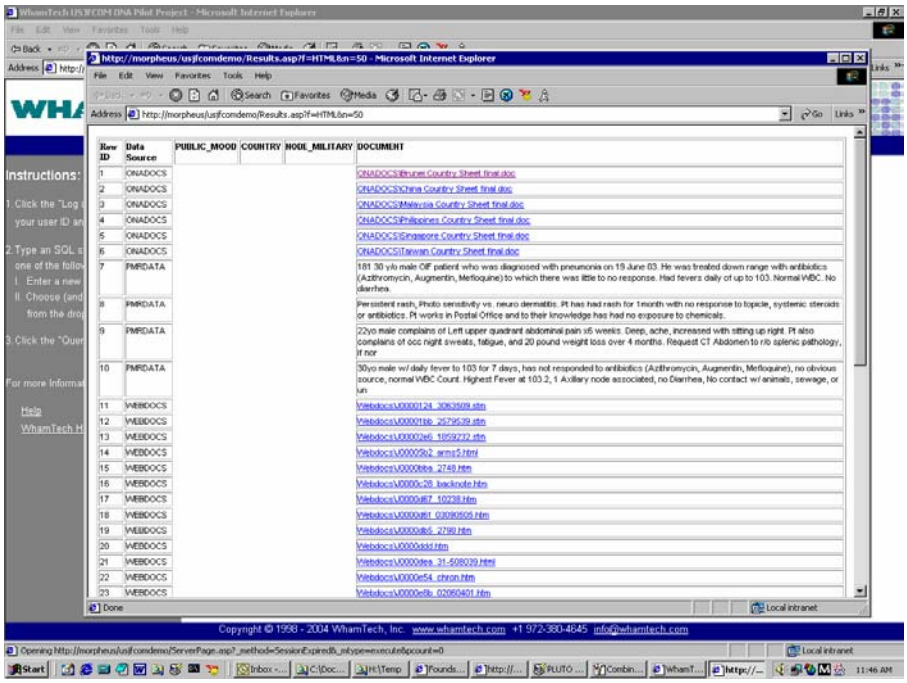
Demo Screen 11: Example of an unstructured text query on using stemming on the word "operation"



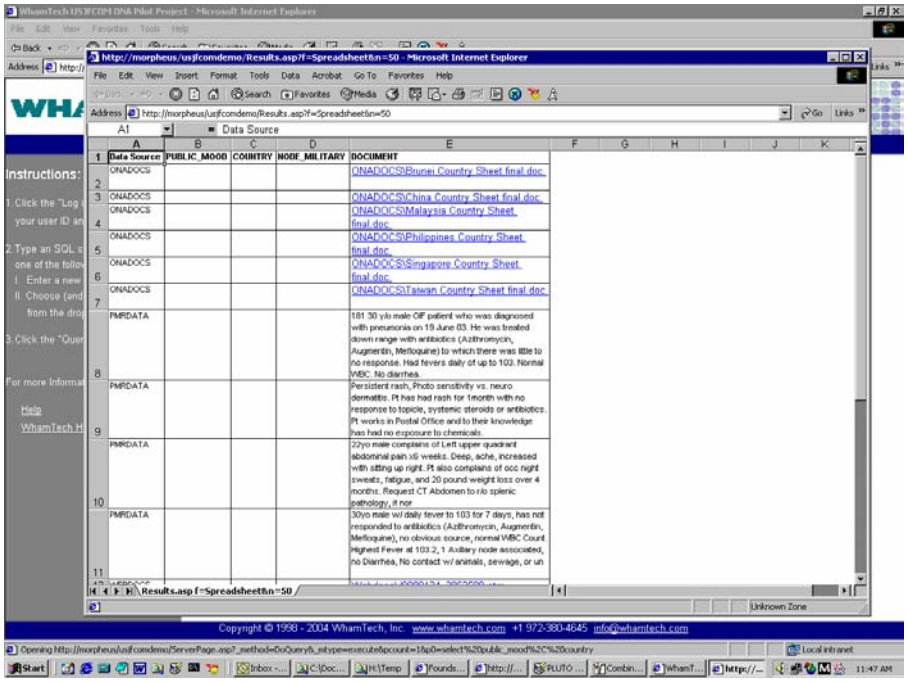
Demo Screen 12: Result-set from the unstructured text query defined in Demo Screen 11



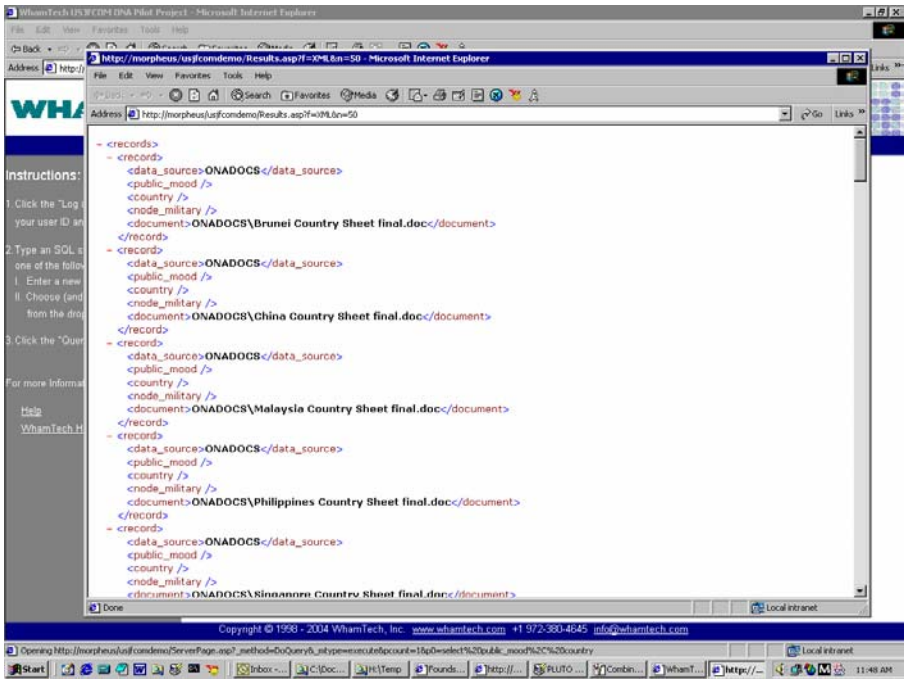
Demo Screen 13: Query for specific relevant metadata fields from the Superschema MyTable using a combined structured data and unstructured text query



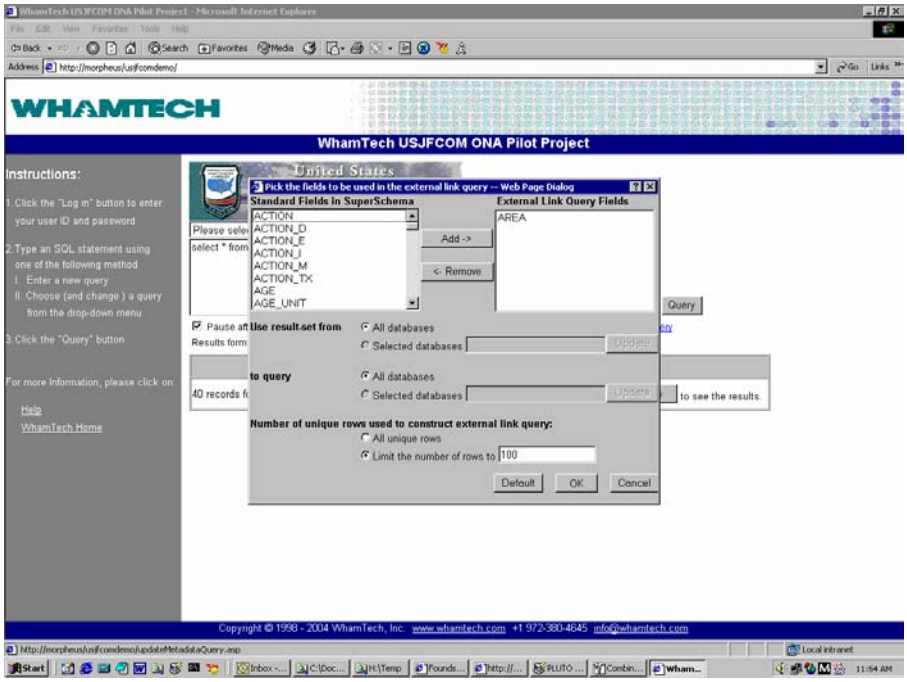
Demo Screen 14: Result-set in HTML format from the query defined in Demo Screen 13. Note the three different data sources shown; more may have been in the result-set



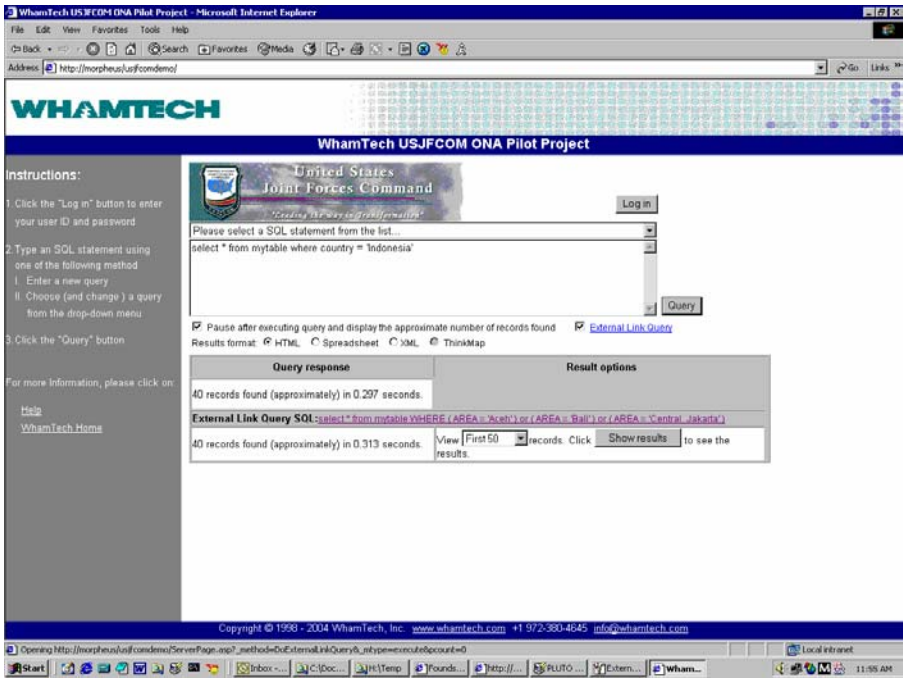
Demo Screen 15: Result-set in Excel format from the query defined in Demo Screen 13. Note that this is still presented in a Web browser



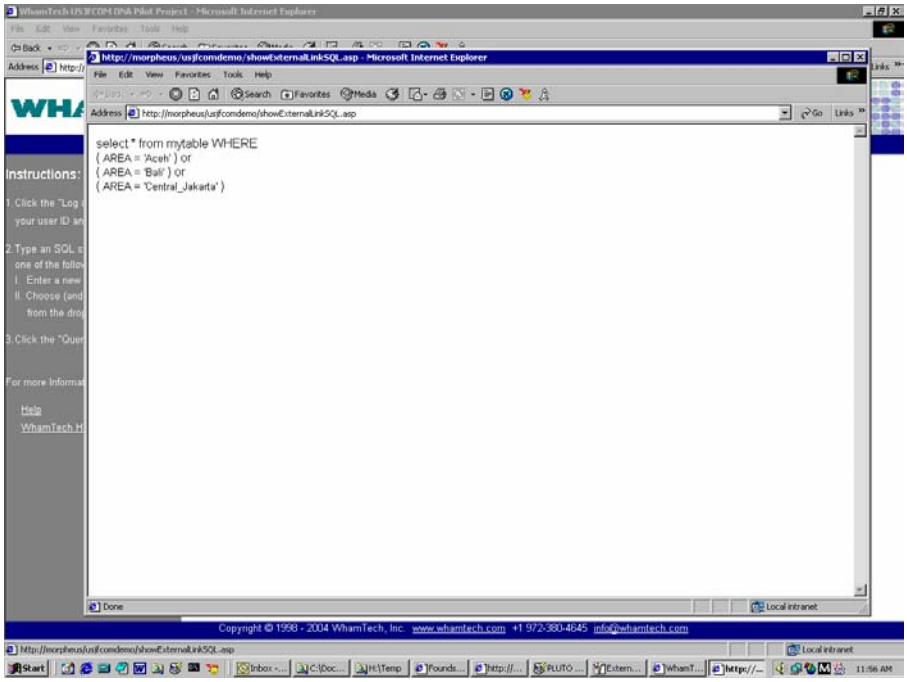
Demo Screen 16: Result-set in XML format from the query defined in Demo Screen 13. Note that this is still presented in a Web browser



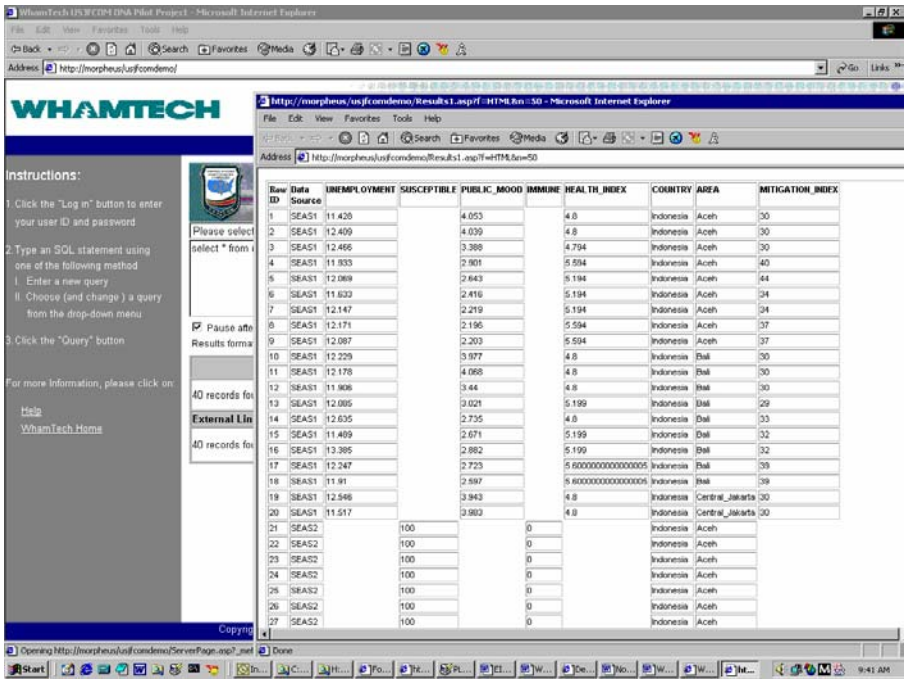
Demo Screen 17: External Link Query metadata field selection screen. In this example, the structured metadata field AREA was used to link across data sources; other fields and Boolean combinations are possible



Demo Screen 18: Results pause screen from the External Link Query defined in Demo Screen 17



Demo Screen 19: External Link Query generated and used to link the AREA field across the data sources. This is performed at the client level; in the future, this will be available at the server level



Demo Screen 20: Result-set from the External Link Query defined in Demo Screen 19. Note that the SEAS data source is the only one that has a structured field called AREA. Other optimizations could allow for AREA being a unstructured text query in addition