ICCRTS 2005
Paper #373

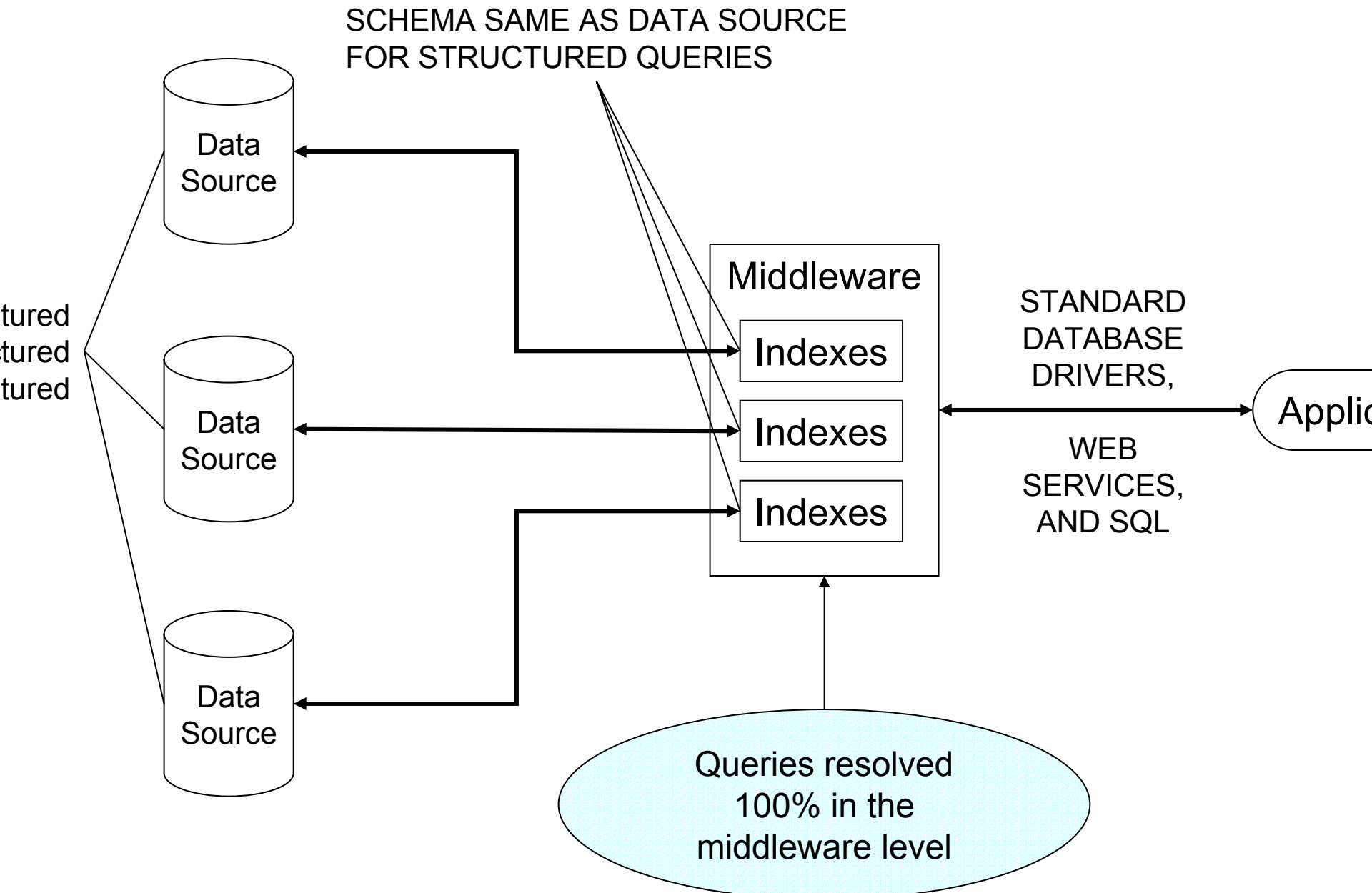
“Applying a Unique Approach in a USJFCOM Joint Experimentation (J9) Rapid Assessment Project for Operational Net Assessment (ONA) Data Integration”

Gavin Robertson, CTO, WhamTech, Inc

Agenda

- What is WhamTech's virtual data integration product (called EIQ Server®)?
- EIQ Server vs. other data integration methods
- USJFCOM Project Description
- Conclusions
- Recommendations
- Questions

EIQ Server


SCHEMA SAME AS DATA SOURCE
FOR STRUCTURED QUERIES
Data Source
Data Source
Data Source
tured
tured
tured
Middleware
Indexes
Indexes
Indexes
STANDARD DATABASE DRIVERS,
WEB SERVICES, AND SQL
Applic
Queries resolved 100% in the middleware level

EIQ Server

OS

Data warehouse performance

- **Find almost 100% of data vs. "up to 50% not found" in federated systems**
- **Indexes/results clean and usable**
- **Complete control over indexing and query processing**
- **Consistent and multiple indexes across disparate data sources**

Data remains at source

No major data and schema transforms

No federated adapters or specialized connectors

Almost any data source

Highly flexible

Security and metadata managed in middleware

Almost no load on data source system

PROS (continued)

- **Data source system/owner unaw of queries**
- **Index monitoring agents feed subscriptions**
- **Fast**
- **Connect to data sources "as us**
  - **Security**
  - **Data updates – two-way**

CONS

- **Establishing index updates**
- **Indexes require storage**

Basic Project Description

The project lasted over three months in 200

- Five data sources selected out of eleven candidates
- DOD XML Metadata Registry was used for metadata
- Indexes were built and maintained external to the data sources
- SQL queries based on metadata including JOINs, range queries, and text search, were executed against the external indexes
- Pointers to result-set data were isolated, and…
- Results retrieved from data sources, integrated and presented in a standard format

Included Data Sources

DS1 – ONA SQL Server relational database and associated Word documents (structured and unstructured – static)

DS3 – TRACES, (stripped) patient medical recor in an Excel spreadsheet (semi-structured – static

DS6 – SEAS PMESII model results of simulating effects of a biological attack in two XML files (semi-structured static)

DS10 – Web documents from ONA-provided new Web sites (unstructured – batch/incremental update)

DS11 – RSS news feeds, including ONA-provide news Web sites (semi-structured – near real-time

Excluded Data Sources

GTN database

ACTD Rosetta

Census data

NGA Fortune Cookie

FBIS Web site

Others

Reasons Data Sources Excluded

Access difficulties

– Even though unclassified, they resided on limited access systems

Owners were reluctant to allow WhamTech to parse and index content

– Not necessarily a copyright issue, but more o a process issue

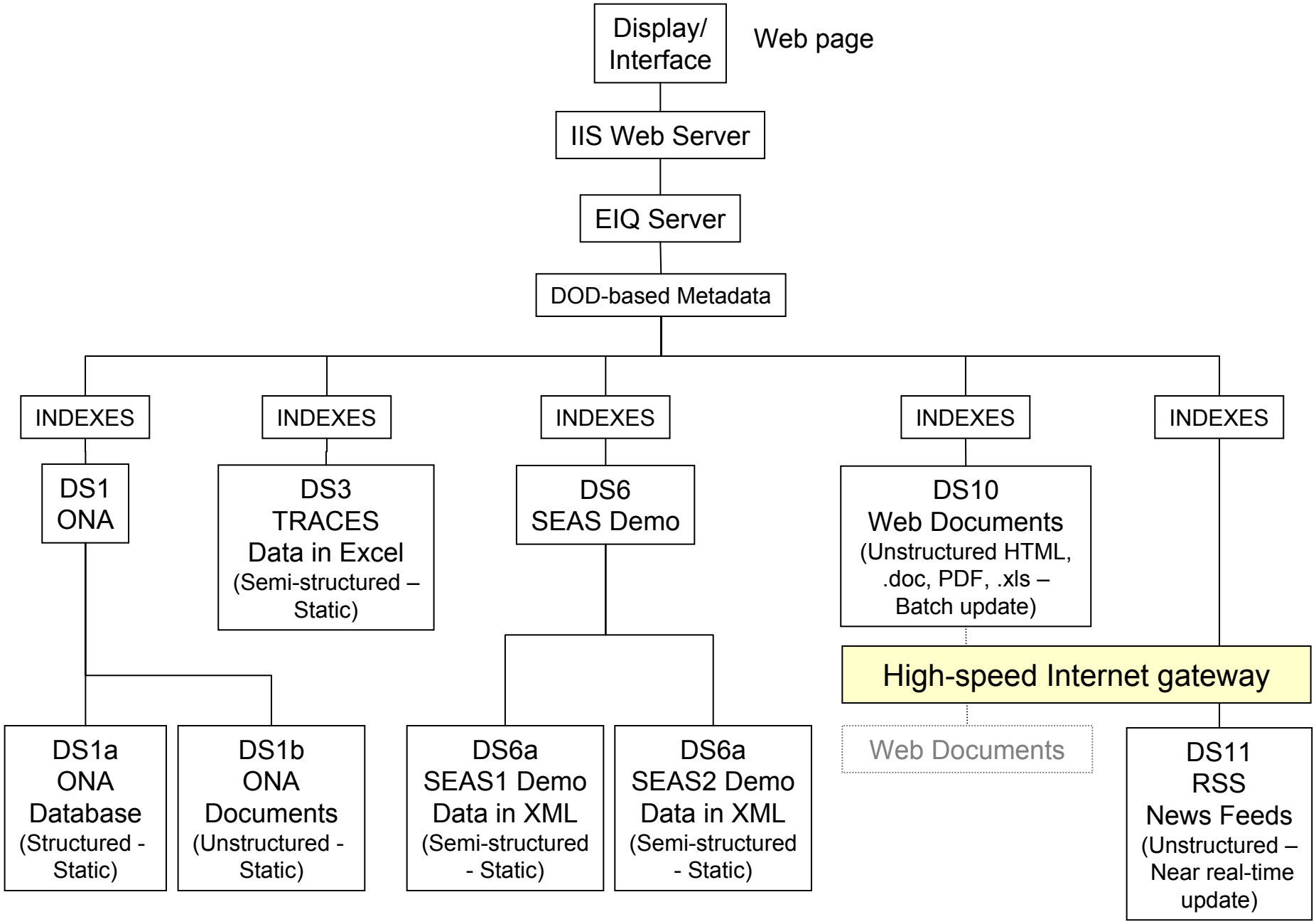
Data was so disparate that there was litt or no commonality

Metadata

Access to the DOD XML Registry

- 30 separate metadata repositories
- Largest was TBD – “To Be Determined” - with over 14,000 data elements
- Of the 30,000+ data elements, a lot of redundancy (overlap)
- Able to use some (~25%) from COAL, GMI, IN PER, and TBD

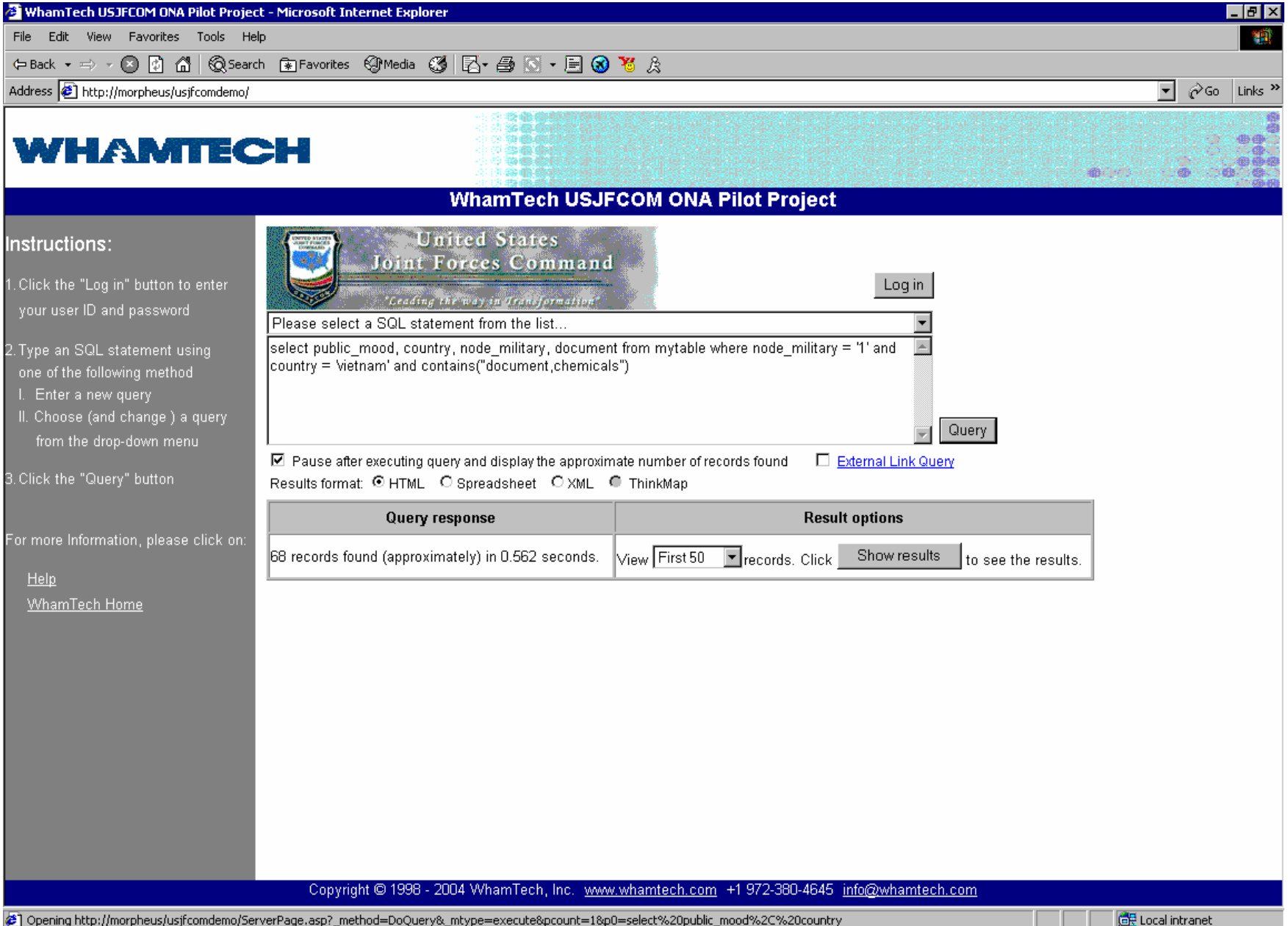
Data Sources and Configuration



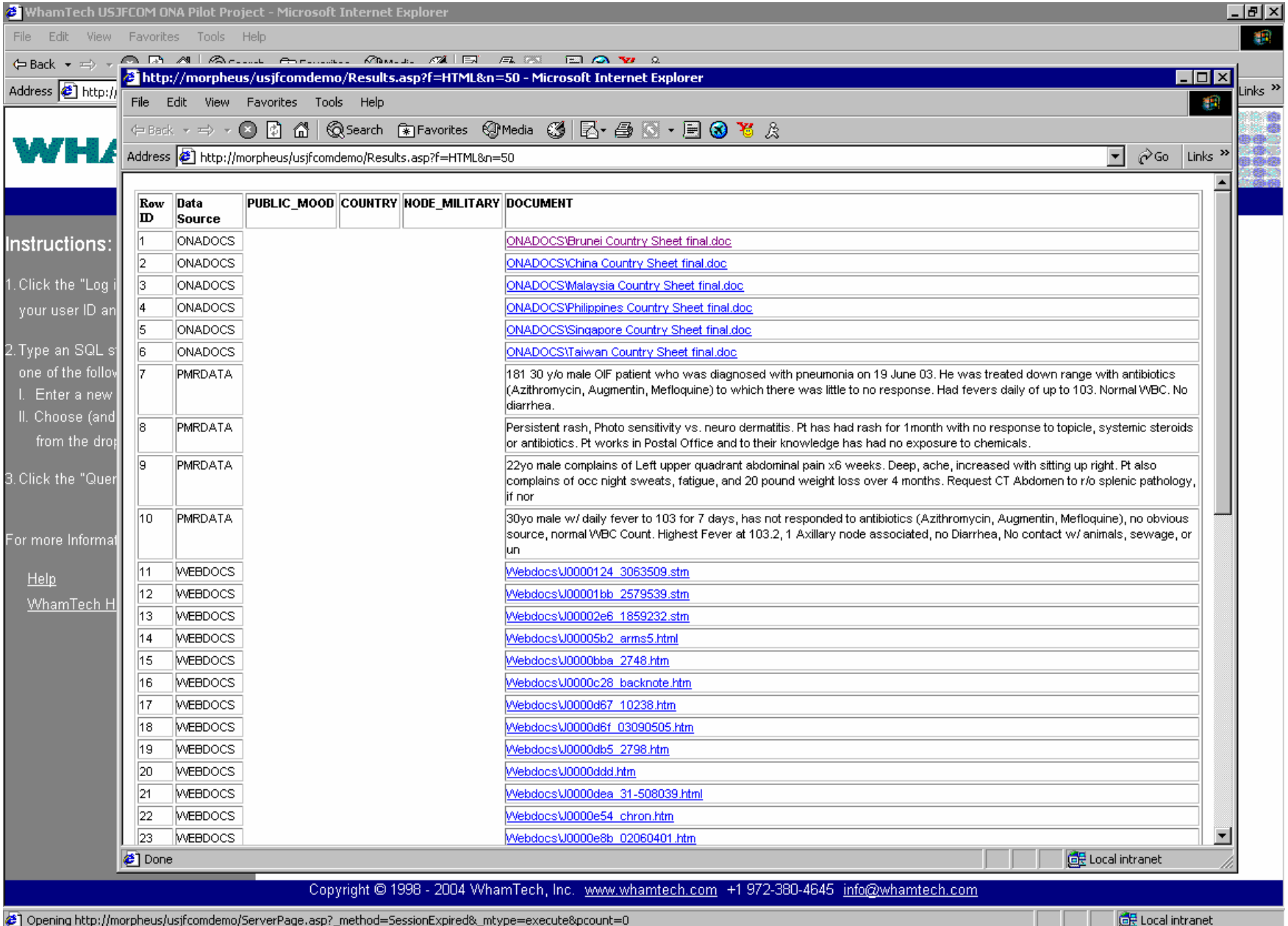
Basic Configuration Process

- Register a data source
- Build an index
- 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 Superschem metadata

and Unstructured Text Search



(other options were Excel and XML)



Row ID	Data Source	PUBLIC_MOOD	COUNTRY	NODE_MILITARY	DOCUMENT
1	ONADOCS				ONADOCS\Brunei Country Sheet final.doc
2	ONADOCS				ONADOCS\China Country Sheet final.doc
3	ONADOCS				ONADOCS\Malaysia Country Sheet final.doc
4	ONADOCS				ONADOCS\Philippines Country Sheet final.doc
5	ONADOCS				ONADOCS\Singapore Country Sheet final.doc
6	ONADOCS				ONADOCS\Taiwan Country Sheet final.doc
7	PMRDATA				181 30 y/o male OIF patient who was diagnosed with pneumonia on 19 June 03. He was treated down range with antibiotics (Azithromycin, Augmentin, Mefloquine) to which there was little to no response. Had fevers daily of up to 103. Normal WBC. No diarrhea.
8	PMRDATA				Persistent rash, Photo sensitivity vs. neuro dermatitis. Pt has had rash for 1month with no response to topicle, systemic steroids or antibiotics. Pt works in Postal Office and to their knowledge has had no exposure to chemicals.
9	PMRDATA				22yo male complains of Left upper quadrant abdominal pain x6 weeks. Deep, ache, increased with sitting up right. Pt also complains of occ night sweats, fatigue, and 20 pound weight loss over 4 months. Request CT Abdomen to r/o splenic pathology, if nor
10	PMRDATA				30yo male w/ daily fever to 103 for 7 days, has not responded to antibiotics (Azithromycin, Augmentin, Mefloquine), no obvious source, normal WBC Count. Highest Fever at 103.2, 1 Axillary node associated, no Diarrhea, No contact w/ animals, sewage, or un
11	WEBDOCS				Webdocs\J0000124_3063509.stm
12	WEBDOCS				Webdocs\J00001bb_2579539.stm
13	WEBDOCS				Webdocs\J00002e6_1859232.stm
14	WEBDOCS				Webdocs\J00005b2_arms5.html
15	WEBDOCS				Webdocs\J0000bba_2748.htm
16	WEBDOCS				Webdocs\J0000c28_backnote.htm
17	WEBDOCS				Webdocs\J0000d67_10238.htm
18	WEBDOCS				Webdocs\J0000d6f_03090505.htm
19	WEBDOCS				Webdocs\J0000db5_2798.htm
20	WEBDOCS				Webdocs\J0000ddd.htm
21	WEBDOCS				Webdocs\J0000dea_31-508039.html
22	WEBDOCS				Webdocs\J0000e54_chron.htm
23	WEBDOCS				Webdocs\J0000e8b_02060401.htm

WhamTech Conclusions

- As an unclassified experiment, access to data sources was restricted or not a high priority
  - Should not be the case in deployment
- Cultural barriers to sharing reflected in a few data source owne responses
- Within DOD, a plethora of metadata dictionaries
  - None for ONA
- More than one metadata dictionary needs to be mapped to sam data
- More than one metadata mapping WITHIN same metadata dictionary
- Need to accommodate variations in so-called standard DOD da
- Could probably use results level indexing instead of data level indexing with ONA and other complex or restricted access data sources
- Novel approach to Excel and XML files, enabling standard drive and SQL access to data as though database tables

-COM J9 Project Alpha Conclusion

- Unique approach to data integration
- Well suited to the ONA process, EBO, JC2, CIE, TIA, HF, and outside DOD in DHS, Intel and law enforcement agencies
- Advantages over data warehousing and federate database approaches
- Despite constraints, able to integrate disparate data sources in real-time
  - Represents an opportunity for ONA analysts to focus analysis than data and information gathering
- Real benefits go beyond time savings…allows th analyst to accomplish more than current processes allow

namTech Recommendations (1 of

Develop a “best of” global common metadata dictionary for data integratio and sharing, e.g. Esperanto

- Don’t force all to adopt – allow application organizations, and countries to continue with own metadata and language
- Map data sources to it
- Map applications to it

Need for ONA metadata dictionary and/or terms

namTech Recommendations (2 of

In an integrated ONA system:

- ONA database as source of query/search terms
- Other systems used as source for the ONA database
- Entity extraction extremely valuable to ONA
- Other KM tools such as semantic reasoning, categorization, a summarization
- Closer to near real-time Assessment - > Planning -> Executio > Assessment loop (EBO)
- Real-time, interactive visualization could add significant value ONA
- Closer tie-in to business process management
- Multiple ONA systems could be integrated and shared at a higher level
- The communities of interest (COI) approach would seem to le itself to ONA and EIQ Servers running ONA
- Multiple COI systems could be integrated and shared at a high level

COM J9 Project Alpha Recommendati

- Novel nature of EIQ Server warrants furth investigation and integration
- EIQ Server approach for other areas than ONA
- Assistant Secretary for Defense for Netwo and Information Integration to include EIQ Server as part of HF and a future Quantur Leap proof-of-concept experiment
- WhamTech seek accreditation for EIQ Server use with classified data sources

Future Plans

- Included in several federal and state agency, and commercial project proposals
- Build on existing implementations as a turbo charger fo RDBMSs and as a much-improved adapter in federate information sharing systems
- Aim to allow almost any application to work with almos any data source
  - Universal metadata management
  - Universal interoperability
  - Ultimate goal: **universal semantic interoperability**
- Reviewing inclusion of Latent Semantic Indexing (LSI)
- Improved entity extraction
- Link Indexes
  - Direct and indirect link analysis in middleware

Acknowledgements

r. Russell Richards, Dr. Kevin Brandt,
Paul (Tom) Fernan and Christian Grar
USJFCOM, J9, Project Alpha

Questions?

avin Robertson
TO & Senior VP
/hamTech, Inc.
450 Sojourn Dr., Suite 200
ddison, TX 75001, USA
l (972) 380-4645 x223
avin.robertson@whamtech.com
ww.whamtech.com

Backup Slides

Different Approach to Metadata Management

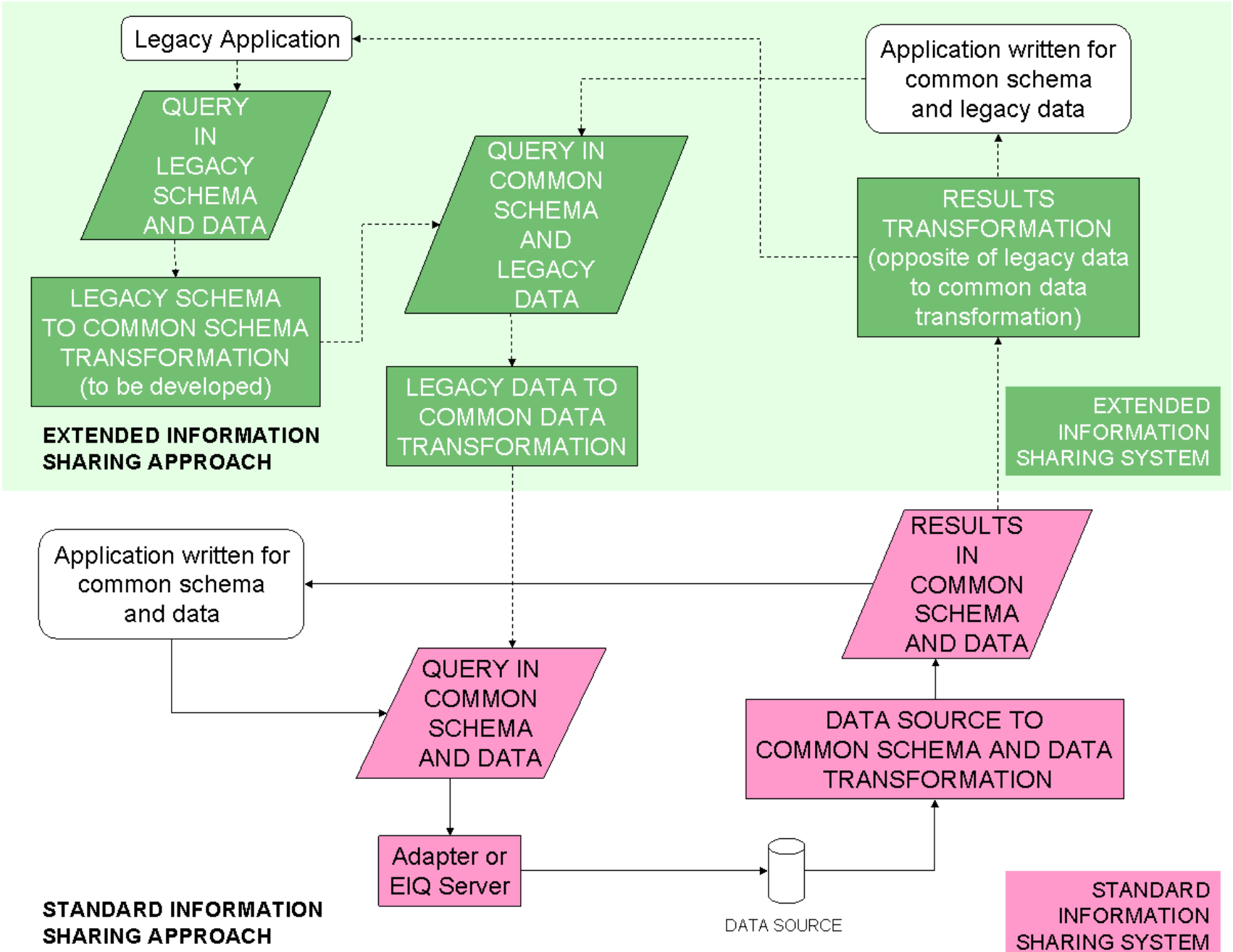
Index Updates

Current EIQ Server System Architecture

Future EIQ Server System Architectur

Back to Questions?

erent Approach to Metadata Managem

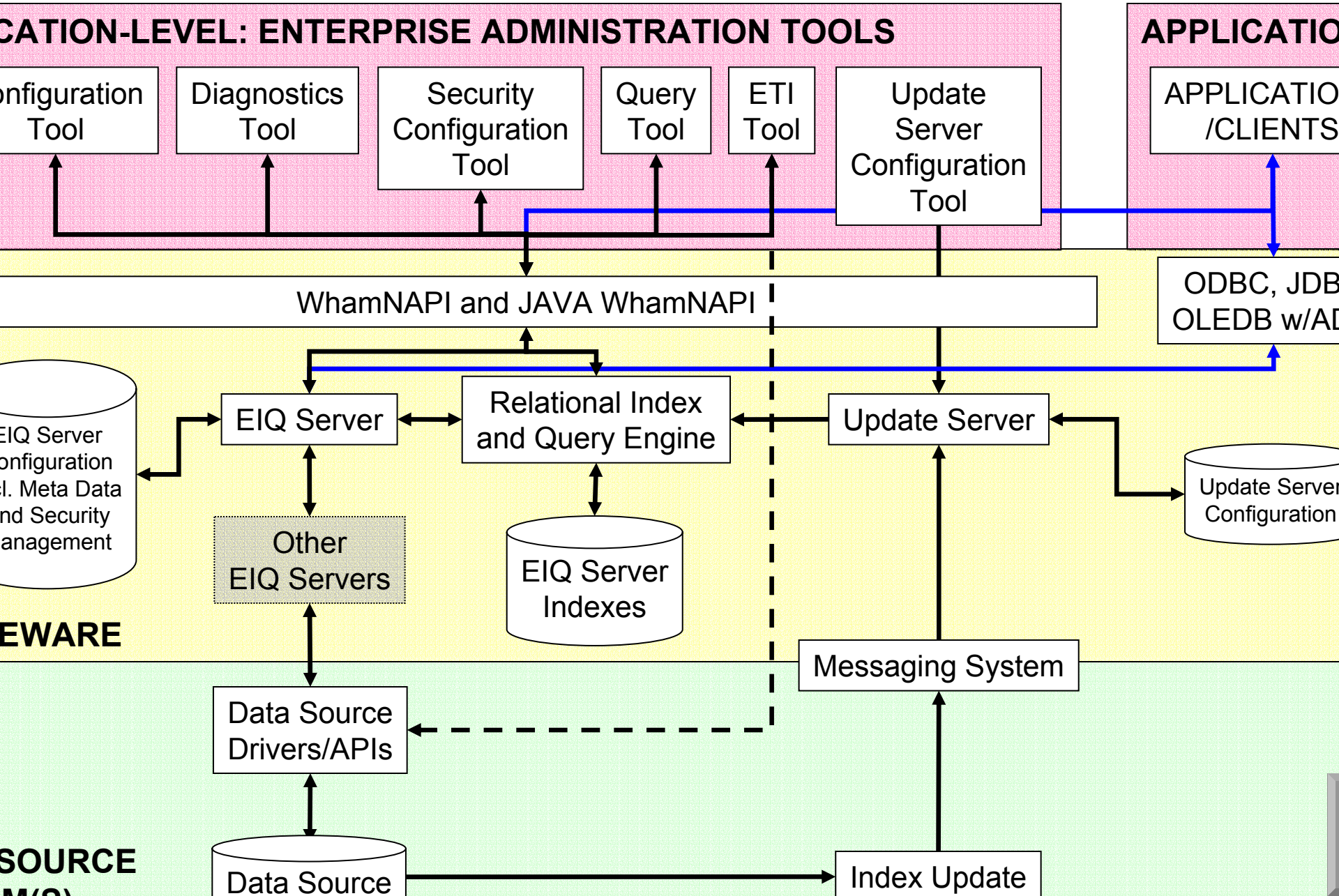
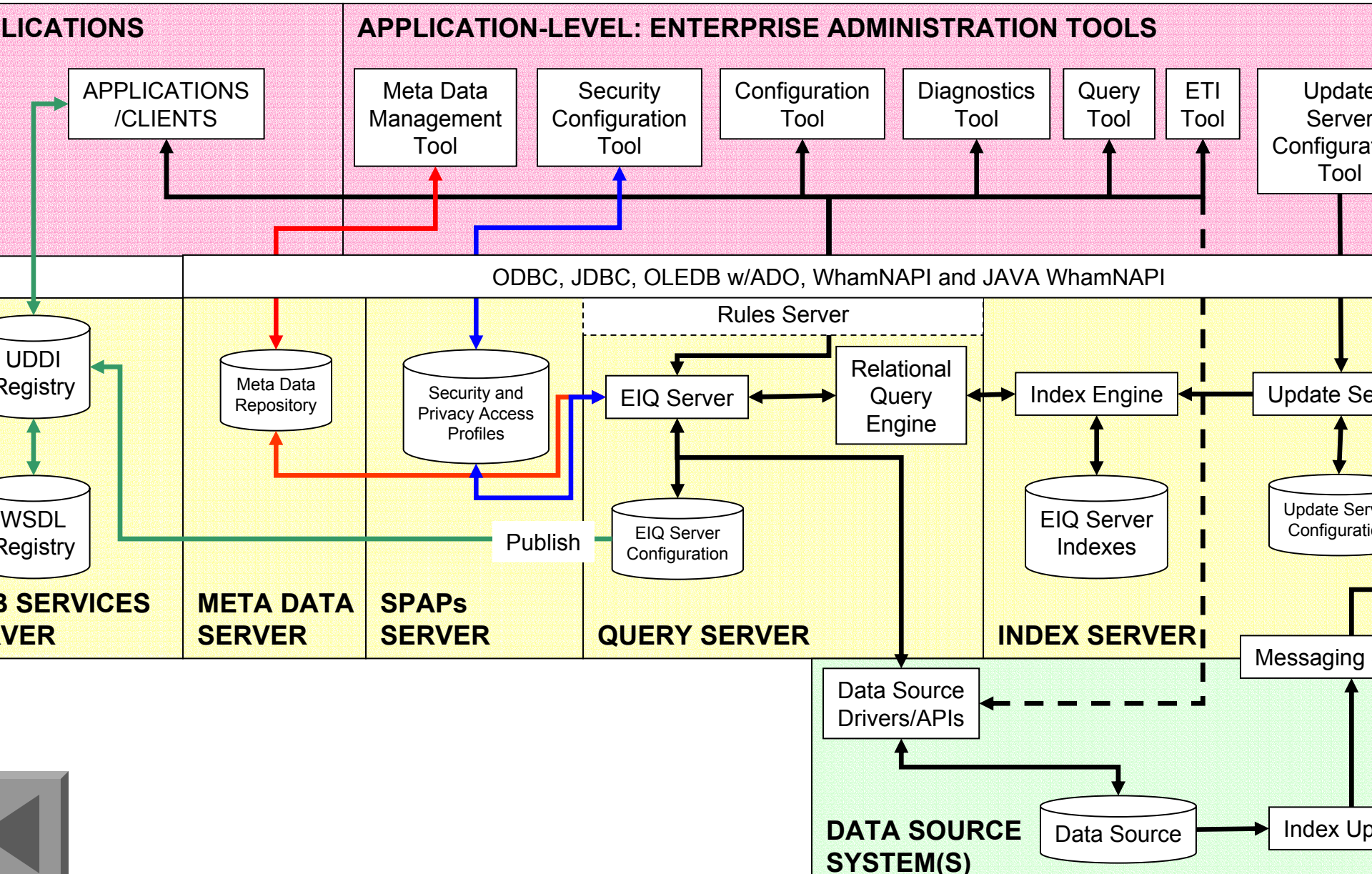


Index Updates

At least seven methods:

- Data level indexing:
  - Batch
    - Complete refresh
  - Incremental
    - Batch updates
  - Transaction or change logs
    - Usually not on data source system
  - Triggers
    - Usually install on data source system
  - Message Queues
    - Tap into
  - Existing replication/backup software
    - Use target as source
- Results level indexing
  - Update rate and route depends on system

rrent EIQ Server System Architect



ture EIQ Server System Architectu